STADLER SYMP. Vol. 12 (1980) University of Missouri, Columbia 83

MOLECULAR EVOLUTION IN PAPOVA VIRUSES AND THEIR HOST SPECIES, AND IN BACTERIOPHAGES

(evolution, nucleotide substitution, papova viruses)

TAKEO MARUYAMA and EIICHI SOEDA

National Institute of Genetics
Mishima, 411 JAPAN

SUMMARY

Comparing homologous genes of three papova viral genomes, we attempt to show the very close relative phylogeny among the viral species and their host species, and therefore the viral species seem to have evolved with their host organisms (SOEDA et al. 1980). Additionally, the DNA-sequence data of bacteriophages ϕX174 and G4 and their overlapping genes will be examined for evolutionary patterns. It will then be made evident that overlapping genes have a quite different substitutional pattern with respect to the position of nucleotides in the codons than do non-overlapping sequences. Namely, in overlapping regions the third positions are usually substituted fastest, followed by the first positions, while the second positions are slowest in each gene, though different genes may have different rates of nucleotide substitution. With overlapping genes, this pattern does not apply, but rather is altered because of an interaction between the substitution rates in the two genes involved in an overlap.

INTRODUCTION

The primary structures of the various proteins were the first source of information for the study of molecular evolution. When the amino acid sequences of homologous genes are compared in taxonomically related species, the similarities and differences are found to be approximately proportional to the closeness of the species. Based on classical taxonomy (FITCH and MARGOLIASH 1967, DAYHOFF 1972), amino-acid-sequence data also enable us to count the number of gene substitutions which have occurred among the species compared. Although the amino-acid-sequence data have provided unique means of studying evolution, the ultimately desirable and most direct data is the nucleotide sequence of genes and genomes. This was not available until the powerful new techniques of DNA sequencing were developed.

The entire genomes of several viral species have been sequenced, as well as many genes of both prokaryotic and eukaryotic species. Usually, DNA-sequence data provide direct evidence for evolutionary studies. For instance, the discovery of a high evolutionary rate at the third positions of the genetic codons and a high incidence of synonymous substitutions have provided rather decisive validation of the neutrality hypothesis of molecular evolution (KIMURA 1977). Using such DNA-sequence information, we intend to point to several interesting phenomena in the evolution of viral genomes and their host species.

MOLECULAR EVOLUTION OF THE PAPOVA VIRUSES AND THEIR HOSTS

The molecular biology of the three papova viruses, BK, SV40, and polyoma (Py), has been extensively studied (FRIED & GRIFFIN 1976, TOOZE 1973, TAKEMOTO & MULLARKEY 1973, SHAH et al. 1977). Earlier work on the genomic organizations of these species has revealed many similarities, and suggests that these papova viruses have diverged from a common ancestor. However, it was not possible to analyze their relationships in detail, until the nucleotide sequences of the viral genomes had been determined. These viral genomes show striking homologies among corresponding genes (REDDY 1978, SOEDA et al. 1980, YAN & WU 1979, SEIF et al. 1979). There is now little doubt that these viral species have diverged from a common origin.

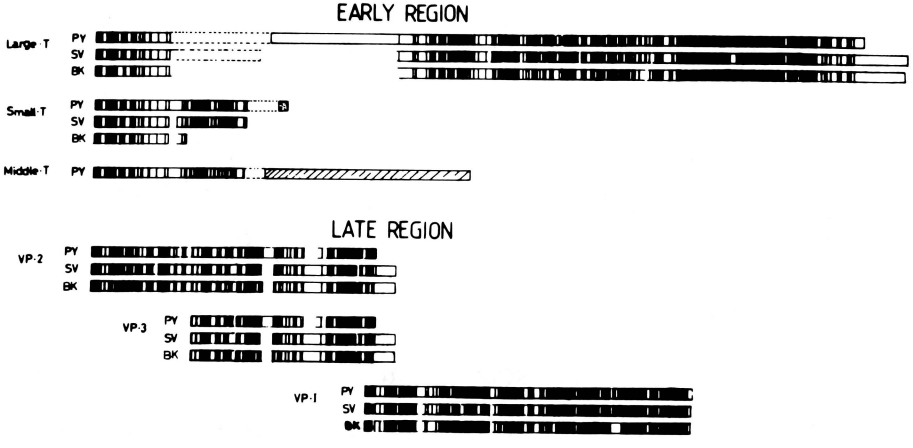

FIGURE 1. A comparison of amino-acid homology in individual genes of Py and SV40, and those of Py and BKV. Black areas represent homologies, and open areas non-homologies. The genes have been linearized about their respective origins of replication, and divided into early and late regions (SOEDA et al. 1980).

Each of these viral genomes consists of two coding regions called the early and late regions, and there are a total of five common genes. The polyoma genome appears to have one additional gene designated as middle T in the early region. A graphic presenta-

tion of homologies in terms of amino acid sequence among genomes is given in Figure 1, where each of the five genes common to the three species is compared. In the figure the homology is viewed from the polyoma genes; the dark areas indicate a strong homology, and the dotted lines indicate splicing regions. The areas where lines are missing indicate inserted gaps.

BKV, SV40, and Py are easily isolated from healthy human, African green monkey, and mouse tissues, respectively (TOOZE 1973, GARDNER et al. 1971). They grow *in vitro* to high titers on cells from these animals. Therefore, we can regard these mammals as their natural hosts. Comparing the DNA sequences of homologous genes, we can calculate linearized relative evolutionary distances and construct a phylogenetic tree for the viral species. For the hosts, the evolutionary time and relationships among species have been well documented from paleontological studies based on the fossil record.

According to paleontology, the order Rodentia, which includes rats, mice, and squirrels, diverged about eighty million years ago from the order Primates, to which various monkeys, the apes, and man belong. Later in mammalian evolution, various forms of Primates have come into existence. In particular, the evolution of the suborder Catarrhini, a well-defined group of Old World Primates including the living monkeys of Asia and Africa, the anthropoid apes, and man, occurred some 35 million years ago in the late Eocene and early Oligocene periods (ROMER 1974, COLBERT 1955). Therefore, applying these paleontological facts to the evolution of the host organism, it seems quite reasonable to suggest that the mouse separated from African green monkey and man some 80 million years ago, and that the green monkey and man diverged from each other 35 million years ago. Figure 2a shows these phylogenetic relationships.

For DNA-sequence comparisons, data were used from four different genes: those coding for t-antigen, T-antigen, and the virus capsid proteins VP_2, VP_3, and VP_1. In the comparisons (made by computer), the corresponding nucleotide sites of DNAs, or amino-acid residues of proteins, were compared and the number of sites or residues at which the two were identical was counted. We then calculated the ratio of the number of sites having identical nucleotides (or amino-acid residues) at corresponding positions to the total number of sites (or residues) examined. We denote this ratio by P_n in the case of the DNA sequence, and by P_a in the case of the (predicted) primary structure of the protein. Thus, P_n is the fraction of the DNA sequence which has the same nucleotides, and P_a the fraction of amino-acid sites having identical residues. We made this calculation for parts of genes for which homology was clear, and therefore a meaningful comparison possible. We have thus excluded from the calculations those parts in which gaps have been inserted in either of the two species being compared. Such gaps appear to exist between Py and SV40 or BKV, but not between SV40 and BKV. With three species to be compared for each of four different genes, there are a total of twelve pairwise comparisons. The results are presented in Table 1.

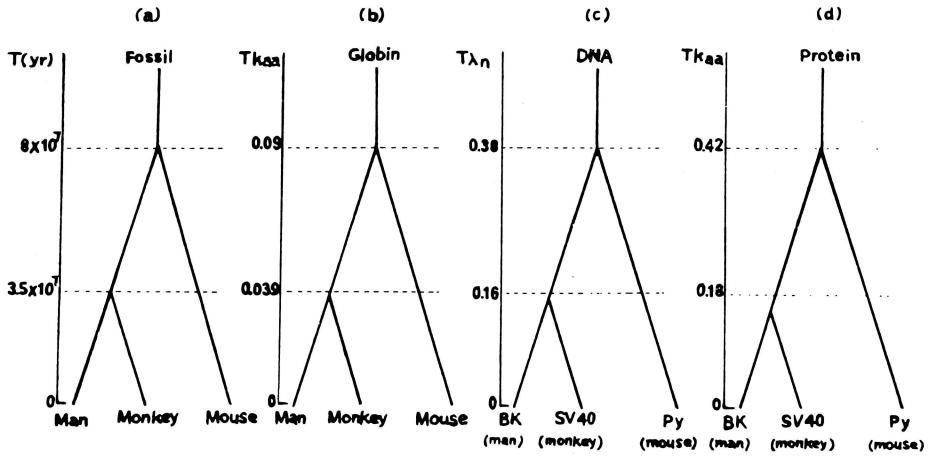

FIGURE 2. Phylogenetic trees of hosts and their viruses. (a) Hosts: based on paleontological evidence; each branch of the tree represents evolutionary divergence time. (b) Homologous globin amino acids: each branch represents the evolutionary distance, measured by kT. (c) Viruses: based on nucleotide sequence; each branch represents the evolutionary distance measured by λT. (d) Viruses: each branch measured by kT. The dotted horizontal lines indicate the relative points where the node should occur if the molecular evolutionary distances among the viruses agreed completely with the branchings of the hosts. Since the dotted lines show only relative locations, we can place one of the two lines at an arbitrary location. We have adjusted the branchings between Py and SV40 (BKV) to the same height of the primates-rodents branching in the figures.

The probability of having identical residues can be translated into a measure of the evolutionary distance between the species compared. The distance measure is linear with respect to time in units of year, and thus when two or more cases are compared, relative evolutionary times can be estimated without knowing the actual rate of evolution. The method, originally developed by ZUCKERKANDL & PAULING (1965), has been extensively used in molecular evolutionary studies (KIMURA 1969, 1977). The principle of the method is as follows: When we look at the evolution of genes at the molecular level, that is, the primary structure of proteins coded for by those genes, the rate at which amino-acid residues become different is constant with time, and the probability of identity between corresponding homologous sites of two species declines exponentially with time. Since there are only a finite number of possible substitutions for both amino acids and nucleotides, this factor must be and has been included in the calculation. If we denote the rate of amino-acid substitution by k, that for nucleotides by λ, and the divergence time by T, we have $P_n = 0.25 + 0.75 \exp(-8\lambda T/3)$ for the nucleotide identity, and $P_a = (1/20) + (19/20)\exp(-40\lambda T/19)$ for the amino-acid-sequence identity. Since P_n and P_a are quantities determined

TABLE 1. Comparison of nucleotide and amino acid sequence among the three virus species BKV, SV40, and Py.

Compared species and genes		Probability of identity		Linearized evolutionary distance; divergence time (T) times substitution rate (λ or k)		No. of nucleotides examined
		P_n DNA	P_a Protein	$2\lambda T$ DNA	$2kT$ Protein	
t-antigen	SV40/Py	0.453	0.306	0.98 (1)	1.25 (1)	519
	BK/Py	0.470	0.322	0.92 (0.94)	1.19 (0.95)	513
	BK/SV40	0.709	0.709	0.37 (0.38)	0.35 (0.28)	516
T-antigen	SV40/Py	0.501	0.423	0.82 (1)	0.89 (1)	1881
	BK/Py	0.425	0.407	1.09 (1.33)	0.93 (1.05)	1380
	BK/SV40	0.721	0.753	0.35 (0.43)	0.29 (0.32)	1535
VP_1	SV40/Py	0.574	0.564	0.63 (1)	0.56 (1)	1080
	BK/Py	0.586	0.532	0.60 (0.95)	0.64 (1.19)	1077
	BK/SV40	0.768	0.818	0.28 (0.44)	0.20 (0.36)	1086
VP_2, VP_3	SV40/Py	0.501	0.434	0.82 (1)	0.86 (1)	885
	BK/Py	0.469	0.341	0.92 (1.13)	1.12 (1.31)	510
	BK/SV40	0.764	0.758	0.28 (0.35)	0.28 (0.33)	534

from the observed data, substituting them into these formulae gives values of λT for the nucleotide comparison and of kT for the amino-acid comparison (Table 1). If the divergence time is known, the rates λ and k can be determined explicitly. For the time being, however, it is important to realize that values of λT and kT, though the products of two quantities (rate and time), are of a linear nature and can be directly compared, while the values of P_n and P_a cannot.

Each comparison gives one value of the evolutionary distance. We have made a total of twelve such comparisons for three species. Although the evolutionary distances calculated from the data did not vary much from gene to gene, we have calculated a weighted-mean distance between each pair of two species to construct a phylogeny. The result for the DNA sequences is presented in Figure 2c, and that for the amino acid sequences in Figure 2d. A remarkable similarity is evident among the host phylogeny based on the fossil record (Figure 2a) and on the two virus phylogenies

constructed from the genome comparison at the molecular level. We take this agreement of the phylogenies as evidence for the co-evolution of the hosts and viruses. That is, these three viral species diverged from each other simultaneously with the divergence of their host species.

It is necessary, however, to examine whether this hypothesis is consistent with other known facts. In the data presented below, we have found no serious contradictions to our hypothesis:

A phylogeny based on the amino acid substitutions among homologous proteins of different species has proved to be in good agreement with relationships inferred from classical taxonomy and paleontology. Proteins often used in constructing molecular evolutionary phylogenies are cytochrome c and hemoglobins (FITCH & MARGOLIASH 1967, 1970). The sequences of hemoglobin α and β are known for a number of species and, when the homologous chains are compared, identity indices (P_a) are found to be: man-monkey, 0.95; man-mouse, 0.85; and monkey-mouse, 0.83. If these P_a values are converted into the linearized evolutionary distance (2kT), we get: man-monkey, 0.074; man-mouse, 0.165; and monkey-mouse, 0.18. From these, a phylogenetic tree of the host species can be constructed. The one based on these data is presented in Figure 2b. With different proteins, although the distance values change, the relative relationships remain approximately the same.

Based on the assumption that these viral genomes have diverged with their host organisms, it is possible to estimate the rates at which the genes in these viral genomes are evolving. Namely, assuming the divergence time between Py and SV40 or BKV to be about 8×10^7 years, and that between SV40 and BKV to be about 3.5×10^7 years, we can get an estimate of λ or k from the linearized evolutionary distance discussed above. We have estimated the rates for each nucleotide position of the codons separately, and also the rate of the amino acid substitutions. The results are presented in Table 2. As noted by various authors, the rates at the third positions of codons appear to be faster by a factor of two to three, as compared with the other two positions (KIMURA 1977). The average rates for the viral nucleotides were found to be about $4.5 \sim 6.5 \times 10^9$ per year per nucleotide site. This value is somewhat larger than the rates estimated for genes in other mammalian genomes, such as the hemoglobin sequences (DAYHOFF 1972).

The rates for the virus proteins were found to be $4.0 \sim 7.8 \times 10^{-9}$ per year per amino-acid site, which is relatively fast. The rates are known for several proteins, and range from 1×10^{-9} for hemoglobins to an exceptionally slow rate of 6×10^{-12} for histone IV.

EVOLUTIONARY PATTERN IN OVERLAPPING GENES IN BACTERIOPHAGES ϕX174 AND G4

The entire genomes of the bacteriophages ϕX174 and G4 have been sequenced (SANGER et al. 1977, GODSON et al. 1979). The two genomes have the same set of 10 genes, and each pair of corre-

TABLE 2. Estimated evolutionary rates of genes in papova viruses BK, SV40, and Py. The rate is estimated from the linearized evolutionary distance, divided by the divergence time.

Gene	Species compared	Nucleotide position in codons				Amino acid	Number of nucleotides or amino-acid sites
		1st	2nd	3rd	Average		
		$\times 10^{-9}$	$\times 10^{-9}$	$\times 10^{-9}$	$\times 10^{-9}$	$\times 10^{-9}$	
t-antigen	Py/SV40	5.4	5.0	8.8	6.4	7.8	171
	Py/BK	5.4	4.9	7.3	5.9	7.4	171
	BK/SV40	4.3	3.1	9.3	5.6	5.0	172
T-antigen	Py/SV40	4.9	3.3	8.7	5.6	5.5	627
	Py/BK	4.8	3.7	8.4	5.6	5.8	460
	BK/SV40	3.1	2.2	10.0	5.1	4.1	512
VP_1	Py/SV40	3.0	2.2	9.4	4.7	3.6	360
	Py/BK	2.6	2.4	7.3	4.1	3.4	359
	BK/SV40	2.9	1.8	8.6	4.4	2.8	362
VP_2, VP_3	Py/SV40	5.1	3.0	9.5	5.9	5.4	295
	Py/BK	5.4	3.7	10.5	6.5	7.0	170
	BK/SV40	3.3	1.4	8.7	4.5	4.0	178

sponding genes shows strong homology, indicating a common origin of the two genomes. GODSON et al. (1979) have provided some comparison of the two genomes, and have shown that there are considerable differences in the evolutionary rates among different genes.

Using the nucleotide-sequence data of ϕX174 and G4, we examined the evolutionary rate of each codon position in these 10 genes. The results are presented in Tables 3 and 4. As is also noted by others, we found that the third positions of codons appear to be the fastest evolving sites among the three positions. However, the three overlapping genes, *B*, *K*, and *E* are totally different in this aspect. As illustrated by Figure 2, gene *B* is contained in gene *A* and consists of 363 nucleotides, providing 121 site comparisons for each of the three nucleotide positions. Although the two genes use the same region of DNA, they have different reading frames and therefore code entirely different proteins. The first nucleotide of every codon in *B* is the third position of *A*, and accordingly, the second and third positions of

TABLE 3. Percentages of identical nucleotides at corresponding sites between genomes of ϕX174 and G4, for non-overlapping or underlying genes.

Gene	Position in codons			Number of nucleotides compared at each position
	1st	2nd	3rd	
A	0.783	0.871	0.583	420
C	0.694	0.812	0.494	85
D	0.810	0.928	0.516	153
F	0.708	0.795	0.452	425
H	0.756	0.824	0.432	324
J	0.808	0.731	0.462	26
G	0.497	0.594	0.352	175

B correspond to the first and second positions of *A*, respectively. A similar reading-frame difference exists between genes *D* and *E*, which is contained in *D* and consists of 276 nucleotides. The 1st, 2nd, and 3rd positions of *E* are respectively the 2nd, 3rd, and 1st positions of *D*. The situation in gene *K* is slightly more complicated. This gene consists of 174 nucleotides, and the first 86 sites are in *A*, which the last 72 sites are in *C*. The four nucleotides from the 83rd to 86th are therefore in both *A* and *C*, where all three genes are coded using the codons triply. In every case of these overlapping genes *B*, *E*, and *K*, the rule that the third positions evolve fastest does not hold.

In the case of gene *B*, there are 121 nucleotides for each position which can be compared between ϕX174 and G4. In *B*, the percentage (P) of identical nucleotides at corresponding sites is 0.645, 0.785, and 0.843 for the 1st, 2nd, and 3rd positions, respectively. This is quite contrary to the rule that the 3rd position is the fastest evolving site. Interestingly, the average of P's, the percent difference, of the three positions is 0.758 for *B* and 0.746 for *A*. Thus *A* and *B* are changing at about the same rate. Now we compare the values of P for the 1st, 2nd, and 3rd positions of *B* to P's of the 3rd, 1st, and 2nd positions of *A*, excluding the part which codes for *B*. Then the corresponding P's are quite similar. Namely, the 1st positions of *A*, excluding the part coding for *B*, have P = 234/299 = 0.783, which is about the same as P = 0.785 of the 2nd positions of *B*. Likewise, 2nd and 3rd positions of *A* have the P values 0.880 and 0.560, which are to be compared with the P values of the 3rd and 1st positions of *B*, being 0.843 and 0.645. Similar comparisons can be made between genes *D* and *E*.

In each of these overlapping genes the fastest-evolving po-

TABLE 4. Incidence of identical nucleotides at corresponding sites in overlapping genes *B*, *E*, and *K* of ϕX174 and G4. The values given for genes *A*, *D*, and *C* are the incidence calculated for the regions, excluding the overlapping regions. Integers in parentheses indicate nucleotide positions in codons of *A*, *D*, or *C*.

Gene	Position in codons			Number of nucleotides compared at each position
	1st	2nd	3rd	
B	0.645	0.785	0.843	121
(*A*)	(3) 0.559	(1) 0.783	(2) 0.880	299
E	0.957	0.620	0.848	92
(*D*)	(2) 0.885	(3) 0.361	(1) 0.754	61
K (1-86)	0.897	0.759	0.821	28 or 29
(*A*)	(2) 0.871	(3) 0.583	(1) 0.783	392
K (83-174)	0.586	0.800	0.833	30 or 31
(*C*)	(3) 0.444	(1) 0.630	(2) 0.796	55

sitions are those corresponding to the 3rd positions of the gene in which the overlapping gene is included. Interestingly, however, the evolutionary rates of these nucleotides in the overlapping genes are somewhat slower than the rates observed in the part outside the overlapping regions. Details of comparisons are presented in Table 4. It is clear that the evolutionary patterns with respect to the nucleotide positions in codons in *B*, *E*, and *K* are very much determined by *A*, *D*, and *C*, though the constraints appear to be slightly more stringent in those overlapping regions. This observation seems to support the hypothesis that these overlapping genes *B*, *E*, and *K* came into existence long after the other genes were established. The emergence of an overlapping gene must be difficult, because the second function must rely on DNA sequences already in use by another gene.

PATTERN OF NUCLEOTIDE SUBSTITUTION

While the third positions of codons are changing more rapidly than the first and second positions, the substitutions at the third positions of codons (informational strand of DNA) do not appear to be symmetrical between species. For instance, at

the corresponding nucleotides in Py and SV40, there are 69 cases with T in Py and C in SV40, while there are 121 cases with C in Py and T in SV40. Therefore, substitutions increasing T in SV40 and C in Py appear to have occurred. In fact, if we examine the number of T's and C's at the 3rd positions of codons in Py and SV40, there are 441 T's and 374 C's in Py, and 594 T's and 212 C's in SV40. The difference seems to be very large. We may say that, in the genome of SV40, the number of T's is increased by decreasing the number of C's, whereas the reverse situation holds in Py.

A similar pattern of asymmetrical substitutions seems to exist between the genomes of ϕX174 and G4. In that case, the 3rd positions of ϕX174 is rich in T, but poor in A, as opposed to G4, which is rich in A but poor in T (in the informational strand of DNA). In Table 5, comparisons of the number of different nucleotides found at each position of the codons are presented; Table 6 provides the comparison of all possible substitutions between species.

TABLE 5. Relative frequencies of T, C, A, and G nucleotides at the lst, 2nd, and 3rd positions of codons in genomes of Py, SV40, ϕX174, and G4.

Position	Organism	T	C	A	G
1st	Py	0.174	0.234	0.288	0.304
	SV40	0.209	0.161	0.309	0.321
2nd	Py	0.269	0.218	0.312	0.201
	SV40	0.277	0.215	0.344	0.164
3rd	Py	0.259	0.221	0.295	0.225
	SV40	0.389	0.125	0.286	0.199
aver.	Py	0.234	0.224	0.299	0.243
	SV40	0.292	0.167	0.313	0.228
1st	ϕX174	0.202	0.215	0.268	0.315
	G4	0.200	0.222	0.277	0.301
2nd	ϕX174	0.270	0.251	0.316	0.163
	G4	0.266	0.267	0.326	0.141
3rd	ϕX174	0.432	0.191	0.164	0.213
	G4	0.344	0.274	0.218	0.164
aver.	ϕX174	0.328	0.239	0.272	0.251
	G4	0.270	0.255	0.274	0.202

The asymmetrical substitutions at the 3rd positions of codons are quite puzzling. The 3rd positions are believed to be more tolerant of substitutions than other positions, because of the high rate of synonymous mutations. This seems to imply that

Table 6. Frequencies of all the possible nucleotide combinations found at corresponding sites of the 3rd positions of codons between Py and SV40, and between ϕX174 and G4.

Nucleotide at the 3rd position in Py	Nucleotide at the 3rd position in SV40			
	T	C	A	G
T	190	69	67	57
C	121	55	82	45
A	134	39	162	94
G	94	27	104	115

Nucleotide at the 3rd position in ϕX174	Nucleotide at the 3rd position in G4			
	T	C	A	G
T	435	216	105	55
C	102	199	42	18
A	40	37	178	52
G	63	63	87	186

if these nucleotide compositions are at equilibrium, substitutions should be approximately symmetrical and independent of the mutation rates among different nucleotides.

Amino acid sequences of proteins have been much used in the molecular evolutionary study of homologous genes. It is generally believed that the evolutionary rate inferred from the primary structures of proteins is not the same over the entire region of a gene. The rate can differ considerably, depending upon the part of the gene examined. Most genes have some part which appears to be extremely conservative, and to have remained unchanged for a long time. This, in terms of nucleotide change, implies that nucleotide substitutions may not be independent among different codons, which can be confirmed at the nucleotide level using the sequence data.

For each of the 1st, 2nd, and 3rd positions, we compute the probability of having identical nucleotides. Then the theoretical expectation of identical or different nucleotides in various combinations can be calculated. If p_1, p_2, and p_3 represent respectively the probability of having identical nucleotides at the

1st, 2nd, and 3rd positions, then for example, the probability of having identical nucleotides at the 1st and 2nd positions but different ones at the 3rd is equal to $p_1p_2(1 - p_3)$, and similarly, that of having different nucleotides at all three positions is equal to $(1 - p_1)(1 - p_2)(1 - p_3)$, and so forth. Of course, these theoretical probabilities are based on the assumption that the substitutions occur independently of each other.

In both comparisons of Py/SV40 and of ϕX174/G4, we confirmed that the nucleotide substitutions are not independent. It is rather striking that the cases of having substitution at either the 1st or 2nd position alone are found only half as frequently as predicted by expectations based on the independence hypothesis. On the other hand, the case of having substitutions at all three positions is found three to four times more often than expected. Similarly, the case of having substitutions at the 1st and 2nd positions but the 3rd position being identical is also found about twice as often than expected. However, interestingly, the case of having a substitution at the 3rd position alone is always in good agreement with expectations. Numerical details are presented in Table 7.

TABLE 7. Three examples of frequencies of nucleotide substitutions in different combinations. Symbols o and x indicate respectively a site having identical nucleotides (no substitution), and a site having no identical nucleotides at corresponding sites in the other species. The symbols o or x indicate, from left to right, the 1st, 2nd, and 3rd position in the codons.

		ooo	oox	oxo	xoo	oxx	xox	xxo	xxx
Py/SV40 large T	Observ.	116	134	27	41	47	98	46	118
	Expect.	73.7	127.3	45.1	69.0	77.9	119.0	42.2	72.8
								χ^2 =	87.63
ϕX174/G4 Gene *A*	Observ.	195	111	13	26	10	35	11	21
	Expect.	166.8	119.1	25.1	46.1	18.0	33.0	7.0	5.0
								χ^2 =	77.6
ϕX174/G4 Gene *E*	Observ.	46	8	29	2	5	1	1	0
	Expect.	46.2	8.3	28.4	2.1	5.1	0.4	1.3	1.4
								χ^2 =	1.4

However, this substitution pattern does not appear to apply to the overlapping genes. This is partly because, in these genes, the relative rates of nucleotide substitution at different positions of the codons are controlled by those genes which underlie the overlapping ones. A typical example is the pattern of substitution in gene *E* of ϕX174 and G4. Interestingly, in *E* the substitutions appear to be independent of each other, and the

agreement between the observed values and the expectations is very good (see Table 7).

The 3rd positions of codons have been shown to be changing considerably faster than the other two positions. Is this because the third positions are far more degenerate than the others, or because the 3rd positions per se have a faster-evolving property? Although the former seems to be the case, it is worth testing. This may be possible using the codons for the amino acids leucine and serine. Leucine has six degenerate codons, TTA, TTG, and CTX, where X indicates any nucleotide of the four. The 1st positions of -TA and -TG are degenerate for T and C, the 3rd positions of TT- are degenerate for A and G, and the 3rd positions of CT- are completely degenerate. Therefore, if all these six codons are equally in use, and if mutation rates are the same for each of the twelve different combinations of nucleotides, we should expect synonymous substitutions at the 1st and 3rd positions in a 1 to 2.25 ratio. Namely, we should find about four substitutions at the 1st positions for every nine at the 3rd positions among the synonymous substitutions for leucine sites. In fact, in the data of ϕX174 and G4, we find 27 synonymous substitutions at the 1st positions and 63 such cases at the 3rd positions for the leucine sites. Therefore the observed ratio is 1: 2.33, which is in good agreement with the expectation, based on the assumption that every position is equally substitutable, if synonymous.

A similar kind of codon degeneracy exists for serine. In this case, all three positions have degeneracy. Namely, TCX, AGT, and AGC code for serine. If these codons are equally frequent, we should expect a 1:1: 2.25 ratio for the 1st, 2nd, and 3rd positions. But in this case, codons AGT and AGC are used considerably less frequently than the others. In the data of ϕX174 and G4, we find the actual numbers of synonymous substitutions to be 13:8:35, which gives a ratio of 1: 0.6: 2.7. Considering the uneven usage of codons for serine, this may also be in agreement with the expectation.

We interpret these observations as evidence that the 3rd position evolves faster, because of high degeneracy.

ACKNOWLEDGMENTS

We are grateful to Dr. Paul Fuerst for valuable comments, and to the Center for Demographic and Population Genetics at the University of Texas at Houston for their assistance in preparing the manuscript.

LITERATURE CITED

COLBERT, E. H. 1955 Evolution of the Vertebrates. Wiley, New York.

DAYHOFF, M. O., ed. 1972 Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington, D.C.

FIERS, W., R. CONTRERAS, G. HAEGEMAN, R. ROGIERS, A. VAN de VOORDE, H. VAN HEUVERSWYN, J. VAN HERREWEGHE, G. VOLCKAERT and M. YSEBAERT 1978 Complete nucleotide sequence of SV40 DNA. Nature 273:113-120.

FITCH, W. M. and E. MARGOLIASH 1967 Construction of phylogenetic trees. Science 155:279-284.

FITCH, W. M. and E. MARGOLIASH 1970 The usefulness of amino acid and nucleotide sequences in evolutionary studies. Evol. Biol. 4:67-109.

FRIED, M. and B. E. GRIFFIN 1977 Organization of the genomes of polyoma virus and SV40. Adv. Cancer Res. 24:67-113.

GARDNER, S. D., A. M. FIELD, D. V. COLEMAN and B. HULME 1971 New human papovavirus (B.K.) isolated from urine after renal transplantation. Lancet 1971, 1(7712):1253-1257.

GODSON, G. N., B. G. BARRELL, R. STADEN and J. C. FIDDES 1978 Nucleotide sequence of bacteriophage G4 DNA. Nature 276: 236-247.

KIMURA, M. 1969 The rate of molecular evolution considered from the standpoint of population genetics. Proc. Nat. Acad. Sci. U.S. 63:1181-1188.

KIMURA, M. 1977 Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267: 275-276.

REDDY, V. B., B. THIMMAPPAYA, R. DHAR, K. N. SUBRAMANIAN, B. S. ZAIN, J. PAN, P. K. GHOSH, M. L. CELMA and S. M. WEISMANN 1978 The genome of simian virus 40. Science 200:494-502.

ROMER, A. S. 1974 Vertebrate Palaeontology, 3rd ed. University of Chicago Press, Chicago.

SANGER, F., G. M. AIR, B. G. BARRELL, N. L. BROWN, A. R. COULSON, J. C. FIDDES, C. A. HUTCHISON III, P. M. SLOCOMBE and M. SMITH 1977 Nucleotide sequence of bacteriophage ϕX174 DNA. Nature 265:687-695.

SEIF, I., G. KHOURY and R. DAHR 1979 The genome of human papovavirus BKV. Cell 18:963-977.

SHAH, K. V., H. L. OZER, H. N. GHAZEY and T. J. KELLY, Jr. 1977 Common structural antigen of papovaviruses of the simian virus 40-polyoma subgroup. J. Virology 21:179-186.

SOEDA, E., J. R. ARRAND, N. SMOLAR, J. E. WALSH and B. E. GRIFFIN 1980 Coding potential and regulatory signal of the polyoma virus genome. Nature 283:445-453.

TAKEMOTO, K. K. and M. F. MULLARKEY 1973 Human papovavirus, BK strain: Biological studies including antigenic relationship to simian virus 40. J. Virology 12:625-631.

TOOZE, J., ed. 1980 Molecular Biology of Tumor Viruses: DNA Tumor Viruses. Pp. 1-73. Cold Spring Harbor Laboratory, New York.

YANG, R. C. A. and R. WU 1979 BK virus DNA: Complete nucleotide sequence of a human tumor virus. Science 206:456-462.

ZUCKERKANDL, B. and L. PAULING 1965 Evolutionary divergence and convergence in proteins. Pp. 97-166. In: Evolving Genes and Proteins (V. BRYSON and H. VOGEL, eds.). Academic Press, New York.