J Mol Evol (1985) 21:221–231

Journal of Molecular Evolution
© Springer-Verlag 1985

Pattern of Nucleotide Substitution and the Extent of Purifying Selection in Retroviruses

Dan Graur

Center for Demographic and Population Genetics, University of Texas Health Science Center at Houston, PO Box 20334, Houston, Texas 77225, USA

Summary. The patterns of point mutation and nucleotide substitution are inferred from nucleotide differences in three coding and two noncoding regions of retroviral genomes. Evidence is presented in favor of the view that the majority of mutations accumulate at the reverse transcription stage. Purifying selection is apparently very weak at the amino acid level, and almost nonexistent between synonymous codons. The pattern of purifying selection obeys the rules previously established in vertebrates [Gojobori T, Li W-H, Graur D (1982) J Mol Evol 18:360–369]; i.e., the magnitude of purifying selection at the amino acid level is negatively correlated with Grantham's [Grantham R (1974) Science 185: 862–864] chemical distances between the amino acids interchanged. We refute Modiano et al.'s [Modiano G, Battistuzzi G, Motulsky AG (1981) Proc Natl Acad Sci USA 78:1110–1114] hypothesis, according to which the pattern of mutation is preadapted to buffer against deleterious mutations. On the contrary, the pattern of mutation reduces the level of conservativeness from that imposed on the amino acid substitution pattern by the structure of the genetic code. The extraordinarily high rate of nucleotide substitution in retroviruses in comparison with that in other organisms is apparently caused by an extremely high rate of mutation coupled with a lack of stringent purifying selection at both the codon and the amino acid levels.

Key words: Retroviruses — Nucleotide substitution — Purifying selection — Reverse transcription — Grantham's chemical distances

Introduction

In a previous study, Gojobori et al. (1982) showed that the difference in the pattern of nucleotide substitution between protein-coding (functional genes) and noncoding regions (pseudogenes) of the vertebrate genome can be explained by assuming that the magnitude of purifying selection at the amino acid level is negatively correlated with Grantham's (1974) chemical distance between the amino acids interchanged. That is, in protein evolution, amino acid substitutions occur mainly between amino acids with similar physico-chemical properties. This observation was later extended to a larger group of eukaryotic genes by Li et al. (1984). The purpose of the study reported here was to establish the patterns and relative rates of nucleotide and amino acid substitution in Retroviridae, and to see whether the rules of purifying selection inferred by Gojobori et al. (1982) are specific to vertebrates, or are applicable to other organisms as well. In this context, retroviruses are particularly useful, since their molecular biology differs from that of vertebrates in many respects. Consequently, finding similar rules of purifying selection at the protein level would be strong evidence that the rules are governed by a universal cause independent of the nature of the organism and are determined only by properties intrinsic to the structure of proteins.

Extensive genetic variation exists among populations of retroviruses (Vogt 1971; Kawai and Hanafusa 1972; Temin 1974; Zarling and Temin 1976), and vast RNA sequence polymorphism is maintained within populations (e.g., see Darlix and Spahr 1983). As a consequence, the genome of a viral species can be described only as an average of a large number of different individual sequences (e.g., see Domingo et al. 1978). RNA genomes are known to evolve 10^4–10^6 times faster than DNA genomes (for a review, see Holland et al. 1982). Gojobori and Yokoyama (1984) have recently shown that in the cellular oncogene c-*mos* the rate of nucleotide substitution is approximately 2×10^{-9} per site per year,

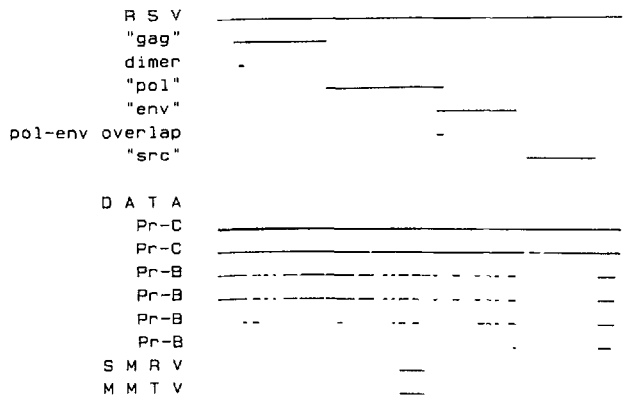

Fig. 1. Schematic representation of the RSV genome and the structure of the available data

whereas in its viral homologue, v-*mos,* this rate is about one million times faster. This high rate of nucleotide substitution presumably reflects (1) an underlying high rate of mutation and (2) lack of stringent purifying selection at the amino acid level. I intend, in the following, to clarify these points, and to investigate into the reasons for this situation in the Retroviridae.

Phylogenetic Relationships

Rous sarcoma virus (RSV) belongs to the avian subgenus of the Oncovirus C genus within the family Retroviridae, the RNA-containing tumor viruses (Davis et al. 1980, pp 1243–1261). The genome of RSV consists of two identical single-stranded 35S RNAs forming a 70S duplex by means of hydrogen bonds. The entire sequence (9312 nucleotides) of the Prague C (Pr-C) strain of RSV has been determined (Schwartz et al. 1983), together with approximately 95% of the sequence of a Pr-C variant present in the population in an amount roughly equal to that of the completely sequenced variant. In addition, fragments that cover approximately 40% of the genome of several spontaneous variants of the RSV Prague B (Pr-B) strain have been sequenced (Darlix and Spahr 1983). Recently portions of the 3′ termini of the *pol* genes in squirrel monkey retrovirus (SMRV), mouse mammary tumor virus (MMTV), and Moloney murine leukemia virus (M-MuLV) were sequenced (Schinnick et al 1981; Redmond and Dickson 1983; Chiu et al. 1984). Initial studies using approximately 540 base pairs within the *pol* gene revealed some small, though statistically significant, degree of sequence similarity among SMRV, MMTV, and RSV (Chiu et al. 1984). With these data, presented schematically in Fig. 1, it is possible to infer the pattern and rate of substitution in various protein-coding and noncoding regions of the RNA genome, and the nature and magnitude of the selective constraints that have resulted in the observed pattern of substitution.

Table 1 shows the nucleotide substitution distances calculated by the method of Jukes and Cantor (1969). Figure 2 shows the UPGMA (Sneath and Sokal 1973) phylogenetic relationship among the viruses. The standard errors of the branching points were calculated by the method of Nei et al. (1984). We can see from the tree that the seven OTUs are clustered by and large according to the intuitively expected phylogeny based on three hierarchical levels of divergence: (1) between viral species (RSV, SMRV, and MMTV); (2) between strains within species (Pr-B and Pr-C); and (3) between variants within strains. The only anomaly observed, which concerns the topological position of one of the RSV Pr-B variants, is resolved if we note that the length of the branch connecting this variant to the rest of RSVs is not significantly different from the length of the next distal branch, connecting the RSV Pr-Bs to the RSV Pr-Cs. The tree in Fig. 2 requires a minimum of 203 nucleotide substitutions (74 synonymous and 129 nonsynonymous). If we change the position of the anomalous RSV Pr-B variant according to the intuitive phylogeny, we obtain a tree that requires 204 substitutions (74 synonymous and 130 nonsynonymous). Consequently, I cannot conclude with any degree of confidence that the anomalous Pr-B strain diverged before the split between Pr-C and Pr-B. Checking most of the 945 possible networks (Cavalli-Sforza and Edwards 1967), I constructed the most parsimonious tree (Fig. 3). This tree requires 199 substitutions (73 synonymous and 126 nonsynonymous), but is anomalous in many respects compared with the phylogeny expected from the hierarchical classification presented above. We see that the distances among the viral species and the RSV strains are, in general, quite large, suggesting either a very long time since divergence or an extraordinarily high rate of nucleotide substitution. The available evidence favors the second alternative.

Pattern and Rate of Nucleotide Substitution

The first step in estimating the pattern of nucleotide substitution is to construct an ancestral sequence and count the number of each type of nucleotide substitution. The requirement is that we have at least three variants; the direction of substitution is then inferred by assuming the most parsimonious pathway. For example, the pentanucleotide sequences spanning positions 2774 to2778 in five RSV variants are CAGTT, CAGTT, TAGTT, CAGGT, and CAGTT, respectively. From these sequences we

Table 1. Matrix of Jukes–Cantor (1969) distances among three species of retroviruses[a]

	MMTV	RSV Pr-C1	RSV Pr-C2	RSV Pr-B1	RSV Pr-B2	RSV Pr-B3
SMTV	0.6982 (0.0545)	0.9482 (0.0762)	0.9287 (0.0743)	1.0238 (0.1052)	0.7969 (0.0879)	0.8652 (0.1158)
MMTV		0.9134 (0.0730)	0.9326 (0.0749)	1.1031 (0.1164)	1.0875 (0.1283)	1.1261 (0.1618)
RSV Pr-C1			0.0113 (0.0046)	0.1773 (0.0250)	0.1373 (0.0242)	0.1343 (0.0287)
RSV Pr-C2				0.1699 (0.0244)	0.1329 (0.0237)	0.1154 (0.0264)
RSV Pr-B1					0.1981 (0.0301)	0.2687 (0.0439)
RSV Pr-B2						0.0730 (0.0205)

[a] Values in parentheses are standard errors

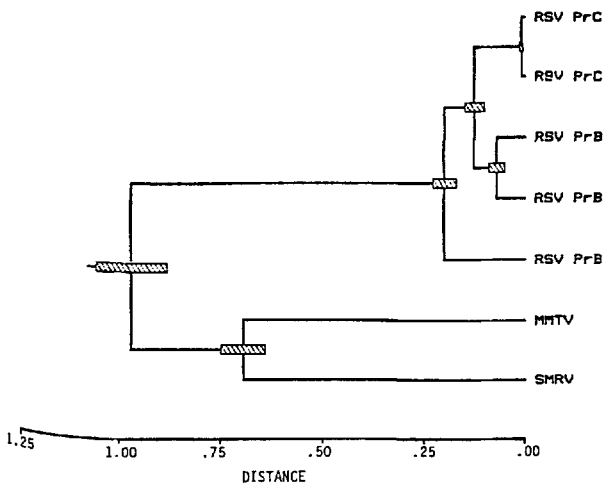

Fig. 2. UPGMA tree of MMTV, SMRV, and two strains of RSV

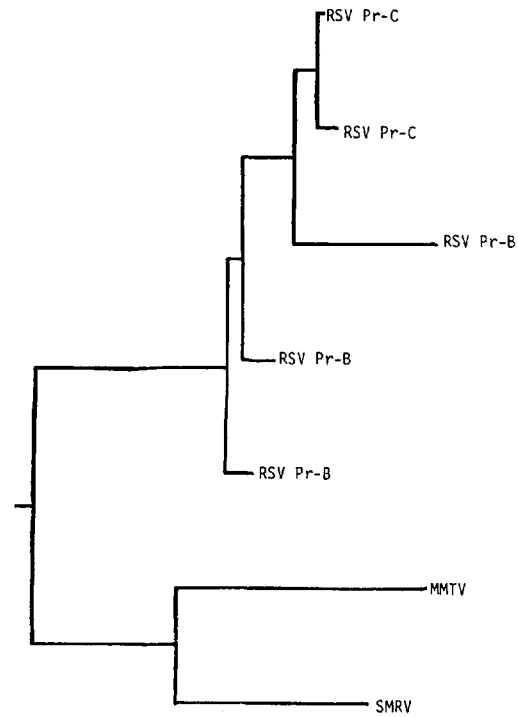

Fig. 3. Parsimonious tree of MMTV, SMRV, and two strains of RSV

can infer the ancestral sequence (CAGTT) and the direction of substitution (i.e., C → T at position 2774 in the third variant and T → G at position 2777 in the fourth variant). In this study I had such data for 3891 nucleotides (42% of the RSV genome). Sixty eight percent of the noncoding regions and 39% of the protein-coding regions of the RSV genome are included in this analysis. The interspecific comparison among RSV, MMTV, and SMRV is based on a stretch of 534 aligned nucleotides within the 3′ terminus of the *pol* gene. The sample size in the interspecific study is, obviously, very small in comparison with the sample size of the RSV study, and I performed the former mainly to check, for longer times of divergence, the conclusions derived from the strain comparisons.

As in Gojobori et al.'s (1982) study, I considered only the untranscribed strand. In 97 sites the ancestral sequence and the direction of nucleotide substitution could not be inferred. To illustrate this possibility, if four sequences to be compared have T, T, C, and C at a particular position, respectively, one cannot decide whether the ancestral sequence had T or C at this position, and whether the substitution was T→C or C→T. One can, however, infer the type of substitution that occurred (T ↔ C). In nine sites, even the type of substitution could not be determined. I excluded deletions or insertions from the present study. In the coding region of the *gag* gene I excluded sites 521–548, since in addition to coding for a protein, this region takes part in forming the dimer linkage structure connecting the two 35S subunits of the RSV genome. The pattern of selection at this site is expected to reflect two

Table 2. Proportions of base substitutions in RSV

Region	A → T	A → C	A → G	T → A	T → C	T → G
			Noncoding regions: all positions			
5′ terminus	4/76	2/76	2/76	1/66	5/66	1/66
Intergene	0/2	0/2	0/2	0/4	0/4	0/4
3′ terminus	4/74	2/74	3/74	7/77	22/77	2/77
Total	8/152	4/152	5/152	8/147	27/147	3/147
			Coding regions: all positions			
gag gene	22/233	8/233	18/233	9/248	26/248	3/248
pol gene	21/342	12/342	13/342	15/329	62/329	8/329
env gene	5/175	6/175	13/175	10/165	29/165	5/165
Total	48/750	26/750	44/750	34/742	117/742	16/742
			Coding regions: first and second codon positions combined			
gag gene	12/165	7/165	9/165	6/158	13/158	2/158
pol gene	16/234	8/234	4/234	11/218	42/218	7/218
env gene	4/121	5/121	7/121	7/107	19/107	3/107
Total	32/520	20/520	20/520	24/483	74/483	12/483
			Coding regions: third codon positions			
gag gene	10/68	1/68	9/68	3/90	13/90	1/90
pol gene	5/108	4/108	9/108	4/111	20/111	1/111
env gene	1/54	1/54	6/54	3/58	10/58	2/58
Total	16/230	6/230	24/230	10/259	43/259	4/259

kinds of constraints, and this might complicate the interpretation of the results.

The direction of substitution was determined for 616 sites. The type of substitution was determined for 706 sites. Table 2 lists the proportions of base substitution (P_{ij}) in coding and noncoding regions of the RSV genome. In coding regions the substitutions are listed according to their positions in the codon (the first and second positions were treated together). Since in each region the rates of substitution are inferred from a different number of variants, I compared the rates of substitution among the different regions of the genome and among the different genes by using the relative rate of substitution per site per lineage. Here I use the word "relative" because I have no knowledge of the times of divergence between either the strains or the viral species. The relative rates of substitution per lineage are also listed in Table 2. From this table I calculated the relative expected proportion of base changes from the i-th type the j-th type nucleotide (f_{ij}) in a random sequence using the equation

$$f_{ij} = \frac{P_{ij}}{\sum_{i} \sum_{j \neq i} P_{ij}} \times 100$$

(Gojobori et al. 1982). The results are listed in Table 3. Cumulative f_{ij} values for transitions are given in brackets in the right upper corner of each matrix.

The rate of substitution in noncoding regions is lower than the rate in coding regions by a factor of approximately 1.5 (Table 2). The rate of substitution in the 3′ terminus is greater than that in the 5′ terminus by a factor of approximately 2. The patterns of substitution for these regions also differ from each other to a considerable extent, suggesting that these two noncoding regions are subject to different selection regimes and stringencies. In both cases transitions occur more frequently than transversions, due mainly to a preponderance of T→C and, to a lesser extent, C→T substitutions. The 5′ terminus exhibits the lowest rate of substitution within the RSV genome, excluding the 17-nucleotide-long intergenic region between the *gag* and *pol* genes, which is perfectly conserved in all four sequenced Pr-B and Pr-C strains. Evidently, the 5′ noncoding region of the genome is subject to strong purifying selection, probably because of its role (among other functions) as a primer-recognition site in reverse transcription (Schwartz et al. 1983). In contrast, the 3′ terminus regions included in this study have very few functions attributed to them, and may thus be subject to less strong purifying selection, as is suggested by their elevated rates of substitution. Yet even the 3′ terminus evolves more slowly than two of the coding regions of the genome (*pol* and *env*) indicating that some function does constrain evolutionary change in this region. The highest rate of substitution among coding regions is seen in the *env* gene, which encodes a glycosylated protein precursor for the envelope of the virion. This protein is largely responsible for

Table 2. Continued

C → A	C → T	C → G	G → A	G → T	G → C	Total	Relative rate of substitution[a]
			Noncoding regions: all positions				
2/89	6/89	0/89	3/125	1/125	1/125	28/356	0.08 (0.0199)
0/5	0/5	0/5	0/6	0/6	0/6	0/17	0.00
3/50	5/50	0/50	5/80	1/80	3/80	57/281	0.20 (0.0378)
5/144	11/144	0/144	8/211	2/211	4/211	85/654	0.13 (0.0280)
			Coding regions: all positions				
9/272	19/272	8/272	15/318	5/318	16/318	158/1071	0.15 (0.0355)
20/378	38/378	15/378	18/385	2/385	21/385	245/1434	0.17 (0.0418)
5/145	14/145	15/145	14/150	5/150	7/150	128/635	0.20 (0.0432)
34/795	71/795	38/795	47/853	12/853	44/853	531/3140	0.17 (0.0400)
			Coding regions: first and second codon positions				
8/200	14/200	5/200	8/198	3/198	6/198	93/721	0.13 (0.0311)
11/263	26/263	8/263	9/243	1/243	15/243	158/958	0.16 (0.0403)
4/92	6/92	9/92	7/105	3/105	7/105	81/425	0.19 (0.0408)
23/555	46/555	22/555	24/546	7/546	28/546	332/2104	0.16 (0.0373)
			Coding regions: third codon positions				
1/72	5/72	3/72	7/120	2/120	10/120	65/350	0.19 (0.0448)
9/115	12/115	7/115	9/142	1/142	6/142	87/476	0.18 (0.0447)
1/53	8/53	6/53	7/45	2/45	0/45	47/210	0.22 (0.0479)
11/240	25/240	16/240	23/307	5/307	16/307	199/1036	0.19 (0.0454)

[a] Numbers in parentheses represent rates of substitution per lineage

determining the host range of the virus, and it exhibits marked variation in nature (for a review, see Vogt 1977). The high levels of substitution may be responsible for this variability.

The rate of substitution for the third nucleotide positions of codons (0.0454) is slightly higher than that for the first and second nucleotide positions combined (0.0373). The third position rate is particularly high for those codons with four-fold redundancies (i.e., TCX, CTX, CCX, CGX, ACX, GTX, GCX, and GGX). The relative substitution rates for these codons are 0.0463, 0.0464, and 0.0491 for *gag*, *pol*, and *env* genes, respectively, with a mean of 0.0471. Assuming that multiple substitutions at a site can be ignored, all substitutions at the third positions of these codons are synonymous, and hence no selection at the protein level is expected to be involved. Moreover, codon usage is by and large random for all amino acids encoded by the four-fold-redundant codons, with the exception of glycine codons (GGX) in *gag* and *pol* genes and valine (GTX) and alanine (GCX) codons in the *gag* gene (Table 4; see also fig. 4 in Schwartz et al. 1983) which are biased. This suggests that for the majority of the amino acids selection barely operates on synonymous codons. In Table 4 we see that the *src* gene, which is not included in this study, has a completely different pattern of codon usage, and is thus thought to have been derived recently from a eukaryotic gene (Schwartz et al. 1983). If we remove the biased codons from our computation, we obtain third position substitution rates of 0.0464, 0.0494, and 0.0480 for *gag*, *pol*, and *env*, respectively, and a mean rate of 0.0482, which is slightly higher than the overall mean of nucleotide substitution for four-fold-redundant positions. The difference, however, is too small to be statistically or biologically significant.

Note that although the rate of nucleotide substitution is consistently higher at the third positions of codons than at the first and second positions, the difference is much smaller than that for eukaryotic genes (Li et al. 1981). It is also smaller than that in viruses other than the Retroviridae, i.e., bacteriophages, papova viruses (Soeda and Maruyama 1982), and, to a lesser extent, influenza viruses (Baez et al. 1980; Krystal et al. 1983; Ortin et al. 1983; N. Saitou, personal communication). This result suggests that purifying selection operating at the amino acid level is much weaker in retroviruses than in any other organisms.

Evaluation of Functional Constraints and their Determinants

Having inferred the patterns of nucleotide substitution in different regions of the genome, we can answer some questions pertaining to amino acid substitution. In particular, we are interested in whether the frequency of substitution correlates with

Table 3. Relative expected substitution frequencies in protein-coding and noncoding regions

Noncoding Regions (All Positions)

5′ Terminus

	A	T	C	G	[56.7]
A	—	15.4	7.7	7.7	30.8
T	4.4	—	22.2	4.4	31.0
C	6.6	19.8	—	0.0	26.4
G	7.0	2.4	2.4	—	11.8
	18.0	37.6	32.3	12.1	

3′ Terminus

	A	T	C	G	[61.3]
A	—	6.8	3.4	5.1	15.3
T	11.4	—	35.9	3.3	50.6
C	7.5	12.5	—	0.0	20.0
G	7.8	1.6	4.7	—	14.1
	26.7	20.9	44.0	8.4	

Total

	A	T	C	G	[60.4]
A	—	9.6	4.8	6.0	20.4
T	9.9	—	33.5	3.7	47.1
C	6.4	14.0	—	0.0	20.4
G	6.9	1.7	3.5	—	12.1
	23.2	25.3	41.8	9.7	

Coding Regions (All Positions)

gag gene

	A	T	C	G	[49.4]
A	—	15.6	5.7	12.8	34.1
T	6.0	—	17.3	2.0	25.3
C	5.5	11.5	—	4.9	21.9
G	7.8	2.6	8.3	—	18.7
	19.3	29.7	31.3	19.7	

pol gene

	A	T	C	G	[54.0]
A	—	8.9	5.1	5.5	19.5
T	6.6	—	27.2	3.5	37.3
C	7.6	14.5	—	5.7	27.8
G	6.8	0.7	7.9	—	15.4
	21.0	24.1	40.2	14.7	

env gene

	A	T	C	G	[54.3]
A	—	3.5	4.2	9.2	16.9
T	7.5	—	21.7	3.7	32.9
C	4.2	11.9	—	12.7	28.8
G	11.5	4.1	5.8	—	21.4
	23.2	19.5	31.7	25.6	

Total

	A	T	C	G	[52.9]
A	—	9.4	5.1	8.6	23.1
T	6.7	—	23.1	3.1	32.9
C	6.3	13.1	—	7.0	26.4
G	8.1	2.0	7.5	—	17.6
	21.1	24.5	35.7	18.7	

Coding Regions (First and Second Codon Positions)

gag gene

	A	T	C	G	[47.2]
A	—	13.9	8.1	10.4	32.4
T	7.3	—	15.7	2.4	25.4
C	7.6	13.4	—	4.8	25.8
G	7.7	2.9	5.8	—	16.4
	22.6	30.2	29.6	17.6	

pol gene

	A	T	C	G	[51.7]
A	—	10.2	5.1	2.6	17.9
T	7.5	—	28.8	4.8	41.1
C	6.3	14.8	—	4.6	25.7
G	5.5	0.6	9.2	—	15.3
	19.3	25.6	43.1	12.0	

env gene

	A	T	C	G	[47.6]
A	—	4.3	5.4	7.5	17.2
T	8.5	—	23.0	3.6	35.1
C	5.6	8.5	—	12.7	26.8
G	8.6	3.7	8.6	—	20.9
	22.7	16.5	37.0	23.8	

Total

	A	T	C	G	[49.9]
A	—	9.7	6.0	6.0	21.7
T	7.8	—	24.0	3.9	35.7
C	6.5	13.0	—	6.2	25.7
G	6.9	2.0	8.0	—	16.9
	21.2	24.7	38.0	16.1	

Coding Regions (Third Codon Positions)

gag gene

	A	T	C	G	[52.8]
A	—	19.2	1.9	17.3	38.4
T	4.3	—	18.8	1.5	24.6
C	1.8	9.1	—	5.4	16.3
G	7.6	2.2	10.9	—	20.7
	13.7	30.5	31.6	24.2	

pol gene

	A	T	C	G	[57.7]
A	—	6.2	5.0	11.1	22.3
T	4.8	—	24.1	1.2	30.1
C	10.5	14.0	—	8.1	32.6
G	8.5	0.9	5.6	—	15.0
	23.8	21.1	34.7	20.4	

env gene

	A	T	C	G	[66.3]
A	—	2.1	2.1	12.5	16.7
T	5.8	—	19.4	3.9	29.1
C	2.1	16.9	—	12.7	31.7
G	17.5	5.0	0.0	—	22.5
	25.4	24.0	21.5	29.1	

Total

	A	T	C	G	[57.7]
A	—	8.9	3.3	13.4	25.6
T	4.9	—	21.3	3.0	28.2
C	5.9	13.4	—	8.5	27.8
G	9.6	2.1	6.7	—	18.4
	20.4	24.4	31.3	23.9	

Total (Four-fold Unbiased Codons)

	A	T	C	G	[66.4]
A	—	11.3	0.0	8.5	19.8
T	2.1	—	6.3	0.0	8.4
C	3.8	25.0	—	5.8	34.6
G	26.6	5.3	5.3	—	37.2
	32.5	41.6	11.6	14.3	

the chemical dissimilarity between the amino acids interchanged. A negative correlation would indicate that substitution occurs mainly between chemically similar amino acids, i.e., that the pattern of substitution is conservative. We shall examine, then, the relative importance of (1) the structure of the genetic code, (2) the pattern of mutation, and (3) purifying selection in determining the degree of conservatism.

Studying amino acid substitution requires large sample sizes, since the analysis is done on a matrix of $(21 \times 20)/2 = 210$ entries. Therefore, the observed relative frequencies of amino acid substitutions (interchanges) within the three genes were pooled. In evaluating the selective constraints at the amino acid level in RSV we encounter one difficulty, no data are available on the intrinsic mutation pattern or rate comparable to the data derived from pseudogenes. I assumed in the folowing that the pattern of substitution at the third positions of unbiased four-fold-degenerate codons reflects the pattern of mutation, since selection at this position was earlier shown to be extremely weak. In the first and second positions the pattern of nucleotide substitution should reflect, in addition to the mutation pattern, constraints imposed by the protein function. To make sure that the differences between the mutation and substitution patterns arise because of constraints on amino acids I used as a control group the noncoding regions, where selection is presumable affected by other, as yet unknown, constraints. In the noncoding regions I artificially divided the sequences into three-letter codons, and combined the three "reading frames." For determining the intrinsic buffering capacity of the "universal" genetic code against radical mutations, I used the frequencies of amino acid substitution due to single-nucleotide substitution that are expected from an equal substitution pattern (Nei 1975, p. 23), and for determining the degree of conservation imposed by the mutational pattern, I used the frequencies of amino acid substitution obtained from the third-position substitution rate of four-fold unbiased codons (Table 3). The equilibrium frequencies of the four nucleotides at this position were calculated by the method of Tajima and Nei (1982). Four thousand pairs of randomly substituted codons were generated.

For each of the four groups of data (coding regions, noncoding regions, and the theoretical expectations from the genetic code and the mutational pattern), I constructed a matrix of amino acid substitutions along the lines of Dayhoff et al. (1972) and determined which amino acids are mutually interchangeable and at what frequencies. The results are presented in Table 5. These frequencies were then correlated with Grantham's (1974) chemical distances for all 210 possible combinations (including synonymous changes). Grantham distances were chosen because they are based on measurable physico-chemical parameters and involve no intuitive classification. Other distances, such as the modified Sneath distances (Doolittle 1979), require clustering of several amino acids into one group, and thus reduce the number of degrees of freedom for the analysis.

Table 4. Deviations from equal usage among four-fold-redundant codons in the Pr-C strains of RSV, as measured by χ^2

Codon	*gag*	*pol*	*env*	*src*
CTX (leu)	4.66	6.11	4.15	90.51**
GTX (val)	10.53*	2.10	2.97	47.37**
TCX (ser)	4.48	7.80	6.93	22.95**
CCX (pro)	0.73	5.41	3.00	57.00**
ACX (thr)	2.90	3.86	6.97	19.70**
GCX (ala)	9.69*	4.62	5.53	17.90**
CGX (arg)	5.54	1.30	3.71	36.88**
GGX (gly)	22.67**	14.73**	1.10	18.78**

Significance of deviation: * $P < 0.05$; ** $P < 0.001$

The first question is whether the genetic code possesses any buffering capacity. In other words, do the allowed substitutions tend to be radical or conservative? Eck (1963) was probably the first to recognize the conservative aspect of the assignment of codons in the genetic code, although his conclusions were based on a slightly erroneous genetic code. Doolittle (1979) concluded on the basis of his modified Sneath distances that "taken as a whole, there is only a slight bias in the code . . . which favors interchanges between similar amino acids" (see also Papentin 1973). The correlation coefficient I obtain between the amino acid subtitution pattern based on equal nucleotide substitutions and Grantham's distances for the case of one substitution per codon is -0.374, and is statistically significant. This result indicates that the structure of the genetic code confers significant protection against radical amino acid replacements.

Next, I examined whether the pattern of mutation introduces a further bias favoring conservative replacements. There are several claims in the literature that this is indeed the case (e.g., see Modiano et al. 1981). My results do not support this claim. When the mutation pattern is taken into account, I detect not an increase, but a slight decrease in the absolute value of the correlation coefficient between the frequencies of exchange and Grantham's chemical distances ($r = -0.368$). The notion of preadaptation of the mutation pattern (i.e., adaptation occurring prior to the action of selection) is not supported. Furthermore, Modiano et al.'s (1981) preadaptation model predicts that the relative frequency of T↔non-T substitutions will be reduced because

Table 5. Observed and expected codon-substitution frequencies (× 1000) in RSV

Noncoding regions (one substitution per codon)

	SER	ARG	LEU	PRO	THR	ALA	VAL	GLY	ILE	PHE	TYR	CYS	HIS	GLN	ASN	LYS	ASP	GLU	MET	TRP
SER	0																			
ARG	0	45																		
LEU	0	0	45																	
PRO	0	0	45	15																
THR	0	0	0	15	30															
ALA	0	0	0	15	30	45														
VAL	0	0	0	0	0	15	15													
GLY	0	15	0	0	0	0	0	45												
ILE	0	0	0	0	106	0	0	0	45											
PHE	0	0	30	0	0	0	0	0	0	0										
TYR	0	0	0	0	0	0	0	0	0	0	15									
CYS	0	30	0	0	0	0	0	0	0	15	0	15								
HIS	0	0	30	0	0	0	0	0	0	0	0	0	0							
GLN	0	0	0	0	0	0	0	0	0	0	0	0	0	0						
ASN	0	0	0	0	0	0	0	0	0	0	15	0	0	0	15					
LYS	0	30	0	0	15	0	0	0	0	0	0	0	0	0	0	0				
ASP	0	0	0	0	0	0	30	30	0	0	0	0	0	0	15	0	45			
GLU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0		
MET	0	0	0	0	15	0	0	0	15	0	0	0	0	0	0	15	0	0	0	
TRP	0	91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Coding regions (one substitution per codon)

	SER	ARG	LEU	PRO	THR	ALA	VAL	GLY	ILE	PHE	TYR	CYS	HIS	GLN	ASN	LYS	ASP	GLU	MET	TRP
SER	42																			
ARG	13	7																		
LEU	15	7	55																	
PRO	33	9	22	40																
THR	20	2	0	4	18															
ALA	2	0	0	9	18	26														
VAL	0	0	15	0	0	26	29													
GLY	4	11	0	0	0	7	7	40												
ILE	0	0	7	0	20	0	20	0	22											
PHE	2	0	11	0	0	0	2	0	2	4										
TYR	0	0	0	0	0	0	0	0	0	2	4									
CYS	4	7	0	0	0	0	0	0	0	0	4	7								
HIS	0	2	4	2	0	0	0	0	0	0	4	0	7							
GLN	0	7	13	2	0	0	0	0	0	0	0	0	7	2						
ASN	2	0	0	0	0	0	0	0	0	0	2	0	0	0	7					
LYS	0	4	0	0	7	0	0	0	2	0	0	0	0	4	2	11				
ASP	0	0	0	0	0	4	4	4	0	0	0	0	2	0	7	0	18			
GLU	0	0	0	0	0	4	9	2	0	0	0	0	0	4	0	4	18	13		
MET	0	0	7	0	4	0	4	0	0	0	0	0	0	0	0	0	0	0	0	
TRP	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0

Theoretical expectations from the genetic code (one substitution per codon)

	SER	ARG	LEU	PRO	THR	ALA	VAL	GLY	ILE	PHE	TYR	CYS	HIS	GLN	ASN	LYS	ASP	GLU	MET	TRP
SER	27																			
ARG	23	34																		
LEU	8	15	34																	
PRO	15	15	15	23																
THR	23	8	0	15	23															
ALA	15	0	0	15	15	23														
VAL	0	0	23	0	0	15	23													
GLY	8	23	0	0	0	15	15	23												
ILE	8	4	15	0	11	0	11	0	11											
PHE	8	0	23	0	0	0	8	0	8	4										
TYR	8	0	0	0	0	0	0	0	0	8	4									
CYS	15	8	0	0	0	0	0	8	0	8	8	4								
HIS	0	8	8	8	0	0	0	0	0	0	8	0	4							
GLN	0	8	8	8	0	0	0	0	0	0	0	0	15	4						
ASN	8	0	0	0	8	0	0	0	8	0	8	0	8	0	4					
LYS	0	8	0	0	8	0	0	0	4	0	0	0	0	8	15	4				
ASP	0	0	0	0	0	8	8	8	0	0	8	0	8	0	8	0	4			
GLU	0	0	0	0	0	8	8	8	0	0	0	0	0	8	0	8	15	4		

Table 5. Continued

	SER	ARG	LEU	PRO	THR	ALA	VAL	GLY	ILE	PHE	TYR	CYS	HIS	GLN	ASN	LYS	ASP	GLU	MET	TRP
MET	0	4	8	0	4	0	4	0	11	0	0	0	0	0	0	4	0	0	0	
TRP	4	8	4	0	0	0	0	4	0	0	0	8	0	0	0	0	0	0	0	0

Theoretical expectations from mutation pattern (one substitution per codon)

	SER	ARG	LEU	PRO	THR	ALA	VAL	GLY	ILE	PHE	TYR	CYS	HIS	GLN	ASN	LYS	ASP	GLU	MET	TRP
SER	44																			
ARG	7	26																		
LEU	16	5	56																	
PRO	35	12	43	25																
THR	17	4	0	4	22															
ALA	4	0	0	12	28	20														
VAL	0	0	12	0	0	26	23													
GLY	13	15	0	0	0	12	3	15												
ILE	2	0	8	0	21	0	23	0	12											
PHE	20	0	42	0	0	0	3	0	8	16										
TYR	3	0	0	0	0	0	0	0	0	7	8									
CYS	11	14	0	0	0	0	0	2	0	2	21	5								
HIS	0	23	7	4	0	0	0	0	0	0	16	0	12							
GLN	0	12	5	2	0	0	0	0	0	0	0	0	6	5						
ASN	9	0	0	0	3	0	0	0	4	0	4	0	2	0	7					
LYS	0	13	0	0	2	0	0	0	1	0	0	0	0	2	6	4				
ASP	0	0	0	0	0	1	4	15	0	0	4	0	6	0	14	0	7			
GLU	0	0	0	0	0	1	3	9	0	0	0	0	0	5	0	5	3	4		
MET	0	1	5	0	6	0	7	0	8	0	0	0	0	0	0	0	0	0	0	
TRP	2	7	2	0	0	0	0	1	0	0	0	5	0	0	0	0	0	0	0	0

organisms are assumed to have adopted the device of lowering the rates of these mutations, which frequently result in radical amino acid changes (using Grantham's distances, we see that T↔non-T substitutions result on the average in an amino acid replacement of magnitude 139, whereas the rest of the substitutions result in a change of 83). In my inferred pattern of mutation the relative frequency of T↔non-T substitutions is 6.25% + 2.08% + 24.98% + 11.34% + 5.32% + 0.00% = 49.97%, as opposed to the random expectation of 50%. T↔non-T substitutions are clearly not less frequent than expected.

As to the purifying-selection pattern, I observe in the protein-coding regions an increase in the absolute value of the correlation coefficient between the frequencies of amino acid substitutions and Grantham's distances ($r = -0.412$). The improvement is quite minor, but since selection is weak we did not expect a dramatic change in the first place. Nevertheless, the rules governing purifying selection at the amino acid level are the same as those in vertebrates (Gojobori et al. 1982), and I conclude again that these rules are determined by the structural requirements of proteins and are independent of the organism. Most replacements occur between amino acids with similar physicochemical properties.

Noncoding regions show exactly the opposite tendency ($r = -0.226$). In these regions the purifying selection, which in magnitude is much stronger than in protein-coding regions, obeys other rules.

Let us now examine the patterns of nuclotide substitution in different retroviral species to see whether or not the above conclusions hold for longer periods of evolutionary times. Note, however, that the results in this part should be regarded as tentative since the problem of multiple substitutions becomes acute in distant comparisons with substitutions exceeding one per site (Table 1), and the amount of data is limited. Statistical analyses similar to the ones discussed above indicate that in the *pol* gene the third position evolves approximately 1.75 times faster (rate of substitution per lineage = 0.290) than the first and second positions (0.166). This implies the existence of weak purifying selection at nondegenerate positions of codons. In the intraspecific comparisons, the frequency of transitions is reduced to 35.7% from 54.3%, supporting the notion that transitions are erased with evolutionary time (Brown et al. 1982). The frequency of synonymous substitutions is also reduced, as expected (Gojobori 1983). The relative frequencies of the six possible bidirectional base substitutions are shown in Table 6. As was done by Gojobori et al. (1982), for each kind of nucleotide substitution, I calculated the expected change in terms of chemical distance (Grantham 1974) between the exchanged amino acids. Based on these distances, we can predict the direction of change (Δf_{ij}) in the relative frequencies of base substitution in protein-coding regions of the genome. For G↔A, C↔G, and C↔A substitutions we expect an increase in frequency in protein-coding regions as compared with the mu-

Table 6. Relative frequencies of base substitution in the pol gene

Substitution	f_{ij}		Δf_{ij}[c]	
	Noncoding[a]	Coding[b]	Obs.	Exp.
C ↔ T	15.1	15.1	=	−
G ↔ A	12.3	16.8	+	+
C ↔ G	17.2	19.1	+	+
C ↔ A	19.4	20.9	+	+
T ↔ G	15.9	14.6	−	−
T ↔ A	20.1	13.5	−	−

[a] Third positions of fourfold unbiased codons
[b] First and second positions of all codons
[c] See text for explanation of Δf_{ij}

tational expectations, because on the average these substitutions result in less drastic amino acid changes than do C↔T, T↔G, and T↔A. These latter changes, in turn, are expected to decrease in frequency in protein-coding regions following the action of purifying selection. The directions of the expected and observed changes in relative frequency of substitution in the interspecific comparisons are listed qualitatively in Table 6. We observe changes in all but the C↔T frequencies, and these are all in the directions predicted by a negative correlation between the magnitude of purifying selection and Grantham's (1974) chemical distances. This supports the notion that the pattern of nucleotide substitution is determined mostly by the chemical relatedness between exchanged amino acids. The pattern of purifying selection is the same as that in vertebrates (Gojobori et al. 1982) and agrees well with the idea that constraints determined by amino acid composition requirements are exerted in protein-coding regions.

Origin of Mutations

One of the most puzzling questions concerning the evolution of Retroviridae is why their rates of nucleotide substitution are so much higher than those of DNA genomes. In Retroviridae mutations can accumulate in any one of three steps in the information flow pathway: (1) reverse transcription (RNA → DNA), (2) replication (DNA → DNA), and (3) transcription (DNA → RNA). Only the first step is absent in DNA genomes, this is also the only step requiring enzymes encoded by the viral genome and RNA as a template. The source of the difference must therefore lie in the reverse transcription process. Indeed, RSV reverse transcriptase has been shown to be a powerful mutagen (e.g., see Battula and Loeb 1974). The present study provides additional evidence that most mutations are introduced during the reverse transcription stage.

The argument runs as follows. In eukaryotic DNA the transition C→T frequently arises from deamination of methylated C residues, which occur almost exclusively at 5′−CG−3′ dinucleotide sites (Coulondre et al. 1978; Razin and Riggs 1980). Interestingly, the majority of ancestral CG sequences in both coding and noncoding regions are mutated in derived sequences (Gojobori et al. 1982). The rate of substitution in CG is much larger than would be expected from the rates of C→non-C and G→non-G separately, resulting in a scarcity of CG in DNA genomes (e.g., see Subak-Sharpe et al. 1966; Nussinov 1981; Wain-Hobson et al. 1981; Konigsberg and Godson 1983). This scarcity has been shown to be closely correlated with the degree of methylation (Gantt and Schneider 1979; Bird 1980). The same is true of RNA genomes lacking the reverse transcription stage (Rothberg and Wimmer 1981). Since in Retroviridae mutations can be introduced both during the replication and transcription stages, when the template (DNA) is heavily methylated, and during the reverse transcription stage, when the template (RNA) is only lightly methylated, and never so in CG positions, one can ascertain the stage at which most mutations occur by comparing the rates of substitution in C and G mononucleotides with the substitution rate in CG dinucleotides.

In RSV no significant increase was detected in the substitution rate in CG dinucleotides. Only 53 out of 171 ancestral CG pairs have changed in at least one derived sequence. The expectation from random substitution is 46.8 changes. The difference is not statistically significant, although the slight increase may arguably be attributed to the very small fraction of mutations occurring at the replication and transcription stages. The results are essentially the same for both coding and noncoding regions, and are homogeneous with respect to region. This observation supports the notion that in RSV most mutations are introduced at the reverse transcription stage. We do find a shortage of CG dinucleotides in the coding regions, but we also find a similar though slightly lesser, shortage in GC dinucleotides, contrary to findings in other organisms. This suggests that the reasons for the shortages are different from those for DNA genomes.

Acknowledgments. I thank Drs. Takashi Gojobori, Wen-Hsiung Li, Masatoshi Nei, J. Clayborne Stephens, and Ciung-I Wu for their help and comments. Naruya Saitou generously shared his unpublished results. This study was supported by research grants NIH GM-20293 and NSF BSF-8315115 to Dr. M. Nei.

References

Baez M, Taussig R, Zazra JJ, Young JF, Palese P, Reisfeld A, Skalka AM (1980) Complete nucleotide sequences of the influenza A/PR/8/34 virus NS gene and comparison with the

NS genes of the A/Udorn/72 and A/FPV/Rostock/34 strains. Nucleic Acids Res 8:5845–5858

Battula N, Loeb LA (1974) The infidelity of avian myeloblastosis virus deoxyribonucleic acid polymerase in polynucleotide replication. J Biol Chem 249:4086–4093

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8:1499–1501

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225–239

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Am J Hum Genet 19: 233–257

Chiu I-M, Callahan R, Tronick SR, Schlom J, Aaronson SA (1984) Major *pol* gene progenitors in the evolution of oncoviruses. Science 223:364–370

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in *Escherichia coli*. Nature 274:775–780

Darlix J-L, Spahr P-F (1983) High spontaneous mutation rate of Rous sarcoma virus demonstrated by direct sequencing of the RNA genome. Nucleic Acids Res 11:5953–5967

Davis BD, Dulbecco R, Eisen HN, Ginsberg HS (1980) Microbiology. Harper & Row, Philadephia

Dayhoff MO, Eck RV, Park CM (1972) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol. 5. National Biomedical Research Foundation, Washington, DC, pp 89–99

Domingo E, Sabo D, Taniguchi T, Weissmann C (1978) Nucleotide sequence heterogeneity of an RNA phage population. Cell 13:735–744

Doolittle RF (1979) Protein evolution. In: Neurath H, Hill RL (eds) The proteins, 3rd edn, vol IV. Academic Press, New York, pp 1–118

Eck RV (1963) Genetic code: emergence of a symmetrical pattern. Science 140:477–481

Gantt R, Schneider WC (1979) Presence of 5 methylcytosine in rat liver mitochondrial and microsomal DNA. In: Usdin E, Borchardt RT, Crevling CR (eds) Transmethylation. Elsevier, New York, pp 465–471

Gojobori T (1983) Codon substitution in evolution and the "saturation" of synonymous changes. Genetics 105:1011–1027

Gojobori T, Li W-H, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369

Gojobori T, Yokoyama S (1984) Rates of nucleotide substitution for cellular and viral oncogenes. Genetics 107:s39

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864

Holland J, Spindler K, Horodyski F, Grabau E, Nicol S, VandePol S (1982) Rapid evolution of RNA genomes. Science 215: 1577–1585

Jukes TH, Cantor CH (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–123

Kawai S, Hanafusa H (1972) Genetic recombination with avian tumor virus. Virology 49:37–44

Konigsberg W, Godson GN (1983) Evidence for use of rare codons in the *dnaG* gene and other regulatory genes of *Escherichia coli*. Proc Natl Acad Sci USA 80:687–691

Krystal M, Buonaguiro JF, Young JF, Palese P (1983) Sequential mutations in the NS genes of influenza virus field strains. J Virol 45:547–554

Li W-H, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237–239

Li W-H, Wu C-I, Luo C-C (1984) Nonrandomness of point mutation as reflected in nucleotide substitution in pseudogenes and its evolutionary implications. J Mol Evol 21: 58–71

Modiano G, Battistuzzi G, Motulsky AG (1981) Non random patterns of codon usage and of nucleotide substitutions in human α- and β-globin genes: an evolutionary strategy reducing the rate of mutations with drastic effects? Proc Natl Acad Sci USA 78:1110–1114

Nei M (1975) Molecular population genetics and evolution. North Holland, Amsterdam

Nei M, Stephens JC, Saitou N (1984) Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol Biol Evol 2:66–85

Nussinov R (1981) Nearest neighbour patterns in the DNA language. In: Balaban M, Sussman JL, Traub W, Yonath A (eds) Structural aspects of recognition and assembly in biological macromolecules. Balaban ISS, Rehovot, Israel, pp 537–549

Ortin J, Martinez C, del Rio L, Davila M, López-Galíndez C, Villanueva N, Domingo E (1983) Evolution of the nucleotide sequence of influenza virus RNA segment 7 during drift of the H3N2 subtype. Gene 23:233-239

Papentin F (1973) A Darwinian evolutionary system. II. Experiments on protein evolution and evolutionary aspects of the genetic code. J Theor Biol 39:417–430

Razin A, Riggs AD (1980) DNA methylation and gene function. Science 210:604–610

Redmond SMS, Dickson C (1983) Sequence and expression of the mouse mammary tumor virus *env* gene. Eur Mol Biol Org J 2:125–131

Rothberg PG, Wimmer E (1981) Mononucleotide and dinucleotide frequencies, and codon usage in poliovirion RNA. Nucleic Acids Res 9:6221–6229

Schinnick TM, Lerner RA, Sutcliffe JG (1981) Nucleotide sequence of Moloney murine leukaemia virus. Nature 293:543–548

Schwartz DE, Tizard R, Gilbert W (1983) Nucleotide sequence of Rous sarcoma virus. Cell 32:853–869

Sneath PHA, Sokal RR (1973) Numerical taxonomy: principles and practice of numerical classification. WH Freeman, San Francisco

Soeda E, Maruyama T (1982) Molecular evolution in papova viruses and in bacteriophages. Adv Biophys 15:1–17

Subak-Sharpe H, Bürk RR, Crawford LV, Morrison JM, Hay J, Kerr HM (1966) An approach to evolutionary relationships of mammalian DNA viruses through analysis of the pattern of nearest neighbor base sequences. Cold Spring Harbor Symp Quant Biol 31:737–748

Tajima F, Nei M (1982) Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. J Mol Evol 18:115–120

Temin HM (1974) The cellular and molecular biology of RNA tumor viruses, especially avian leukosis-sarcoma viruses, and their relatives. Adv Cancer Res 19:47–104

Vogt PK (1971) Spontaneous segregation of nontransforming viruses from cloned sarcoma viruses. Virology 46:939–946

Vogt PK (1977) The genetics of RNA tumor viruses. In: Fraenkel-Conrat H, Wagner RR (eds) Comprehensive virology, vol 9. Plenum Press, New York, pp 341–455

Wain-Hobson S, Nussinov R, Brown RJ, Sussman JL (1981) Preferential codon usage in genes. Gene 13:355–364

Zarling DA, Temin HM (1976) High spontaneous mutation rate of an avian sarcoma virus. J Virol 17:74–84

Received July 16, 1984/Revised November 20, 1984