Review Letter

# VIRAL, PROKARYOTE AND EUKARYOTE GENES CONTRASTED BY mRNA SEQUENCE INDEXES

Richard GRANTHAM

*Laboratoire de Biométrie, Université Claude Bernard – Lyon I, 69621 Villeurbanne, France*

Received 22 August 1978

## 1. Introduction

Comprehensive comparison of genes of different kinds of species is becoming feasible by studying messenger RNA sequences, 29 of which have been published that contain 50 or more completely determined continuous codons. These 29 sequences are analyzed here for several characteristics believed useful for distinguishing molecular strategies of evolution.

Of the three viral genomes entirely sequenced, MS2 is a single-stranded RNA coliphage [1], $\phi$X174 a single-stranded DNA coliphage [2], and SV40 a double-stranded oncogenic primate virus [3]. Use of the codon catalog varies greatly among these viruses. For example, codons of the type NCG (where N is any base) are completely absent in SV40, but relatively abundant in other viruses [1–7]. It was noted early that G tends to be rare as third base of highly degenerate viral codons [8], and this has since been found generally true in MS2, $\phi$X174 and SV40. However, SV40 is the extreme case: for 326 of the 1511 codons in its best-defined translated sequences, opportunity exists for use of an NCG triplet, but each time A, C or U appears as third base. Also, the frequency order of degenerate bases in SV40 codons (C<G<A<U) is the same in each of the four main genes, but greatly different from that in animals or other viruses, as seen below.

I now calculate frequencies of each kind of two-base sequence (doublet), following Russell and Subak-Sharpe [9], to aid in a quantitative differentiation of these viruses and other species. Among the 16 possible base doublets, CG has the lowest frequency by a wide margin in the SV40 genome. The CG doublet is rare, not merely as observed for the NCG codons (that is, in codon positions II–III), but also for the two other combinations of codon position (I–II and III–I). The CG doublet is rare, although to a lesser degree, in the untranslated part of the genome as well, as observed [3]. In this virus, therefore, the CG doublet is discriminated against independently of its potential for amino acid coding (in codon positions I–II, CG codes arginine). The CG doublet is more rare in SV40 than is any doublet in any other genome.

This study may provide indications on viral origins and on the molecular constraints existing during gene and genome evolution. My analysis demonstrates that SV40 differs enormously from other viruses and from the coding nucleic acids of the mammalian cell it parasitizes.

## 2. Results and discussion

The ratio of observed to expected number for each of the 16 doublets is seen in table 1. These data are for the coding regions in the three viral genomes and for each animal mRNA sequence of 50 or more completely determined contiguous codons. The SV40 value of 0.06 for the CG doublet is the lowest entry. In general, CG is the most variable of all doublets.

The ratio of observed frequencies for CG and GC doublets is the first index to be exploited. The CG/GC doublet frequency ratios of the three viral genomes are summarized in table 2 and compared to prokaryote and eukaryote genes sequenced. In MS2 and $\phi$X174, the CG/GC ratio is roughly the same in

Table 1
Nucleotide doublet frequencies in mRNA

| mRNA sample | AU | UA | AA | UU | GU | AC | UG | CA | GA | UC | AG | CU | GG | CC | GC | CG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MS2 [1] | 0.93 | 0.95 | 1.14 | 0.96 | 1.03 | 0.98 | 0.90 | 0.95 | 0.97 | 1.18 | 0.95 | 1.07 | 1.01 | 0.85 | 0.98 | 1.12 |
| $\phi$X174 [2] | 0.86 | 0.78 | 1.34 | 1.15 | 0.89 | 0.96 | 1.10 | 0.98 | 0.98 | 0.91 | 0.91 | 1.07 | 0.91 | 0.90 | 1.25 | 1.02 |
| SV40 [3] | 0.75 | 0.70 | 1.20 | 1.20 | 0.84 | 0.97 | 1.35 | 1.19 | 0.95 | 0.71 | 1.05 | 1.34 | 1.15 | 1.31 | 1.15 | 0.06 |
| *Psammechinus* H2B [11] | 0.63 | 0.49 | 0.96 | 1.25 | 1.41 | 0.81 | 0.91 | 1.29 | 0.91 | 1.28 | 1.67 | 0.89 | 0.84 | 1.05 | 0.91 | 0.66 |
| Chicken ovalbumin [12] | 0.95 | 0.48 | 1.03 | 1.01 | 0.80 | 0.69 | 1.41 | 1.45 | 1.11 | 1.20 | 1.32 | 1.28 | 0.91 | 0.98 | 1.19 | 0.19 |
| Mouse $V\lambda_{II}$ Ig [13] | 0.76 | 0.75 | 0.76 | 0.78 | 0.96 | 1.33 | 1.34 | 1.47 | 1.00 | 1.18 | 1.12 | 1.47 | 1.46 | 0.76 | 0.70 | 0.07 |
| Rat preproinsulin [14] | 0.35 | 0.38 | 1.44 | 0.81 | 0.98 | 1.20 | 1.49 | 1.25 | 0.92 | 0.99 | 1.02 | 1.53 | 1.20 | 0.99 | 0.88 | 0.42 |
| Rat growth hormone [15] | 0.77 | 0.42 | 1.14 | 0.98 | 0.56 | 0.74 | 1.52 | 1.17 | 1.10 | 0.91 | 1.40 | 1.55 | 0.84 | 0.93 | 1.38 | 0.44 |
| Rabbit beta globin [16] | 0.58 | 0.36 | 1.70 | 0.67 | 1.06 | 0.85 | 1.54 | 1.39 | 0.74 | 1.01 | 1.10 | 1.69 | 1.10 | 1.02 | 1.08 | 0.13 |
| Human CS lactogen [17] | 0.94 | 0.61 | 1.03 | 0.81 | 0.63 | 0.94 | 1.32 | 1.07 | 1.13 | 1.12 | 1.09 | 1.46 | 1.09 | 0.90 | 1.06 | 0.58 |
| 7 Animal [11–17] | 0.80 | 0.52 | 1.06 | 0.90 | 0.84 | 0.87 | 1.43 | 1.34 | 1.01 | 1.08 | 1.28 | 1.45 | 1.03 | 0.94 | 1.11 | 0.34 |
| 5 Mammalian [13–17] | 0.71 | 0.54 | 1.09 | 0.81 | 0.82 | 0.99 | 1.48 | 1.29 | 0.99 | 1.03 | 1.21 | 1.57 | 1.08 | 0.91 | 1.08 | 0.36 |

Values are for all nearest neighbor doublets in each sequence. The top three samples are complete translated genomes of several genes each. Initiator and terminator codons are excluded since they are counted as untranslated. The next seven samples are the animal mRNA so far sequenced of at least 50 contiguous codons entirely determined. It is not certain that the mouse $V\lambda_{II}$ sequence is all translated [13]. The ratio of observed to expected number (calculated from individual base frequencies in each codon position) appears under each doublet. For MS2, $\phi$X174 and SV40, doublets in each gene were counted and added to those in the other genes of the same genome. In SV40, values for the CG doublet by codon position combination are: positions I–II, 0.12; II–III, 0.00 and III–I, 0.08. The nine CG doublets present in the coding region of the SV40 genome are scattered among the different genes. In this table (only), the overlap of nucleotides 1425–1535 has been counted twice, however, once for VP1 and once for VP2. A total of 10 CG doublets and 219 GC doublets results by this gene by gene method of counting. Since ratios of observed to expected numbers are involved and the overlap only represent 2.5% of the codons in the four major genes [3,28], the values reported are hardly affected by the counting method. VP3 of SV40 and $\phi$X174 frameshift genes E and K are not included here because they are contained in other coding sequences of the same genomes [2,3,29,30]. The $\phi$X174 genes are thus A, D, F, G, H and J [2]. SV40 genes are VP1, VP2, T and t [3,30,31]. Also determined in [30], the SV40 sequence taken here is from [3]. The bottom two lines are for the combined seven animal and five mammalian mRNAs, respectively

Table 2
Genome and gene characteristics of viruses and hosts

| Genome or mRNA | Total bases | Translated region | | | | | | Untranslated region | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | bases | CG/GC | CG | GC | expected | bases | CG/GC | CG | GC |
| MS2 [1] | 3569 | 3 | 3195 | 1.18 | 245 | 207 | 212.6 | 374 | 0.88 | 22 | 25 |
| $\phi$X174 [2] | 5375 | 6 | 4863 | 0.76 | 246 | 322 | 247.0 | 468 | 0.79 | 19 | 24 |
| SV40 [3] | 5226 | 4 | 4425 | 0.04 | 9 | 213 | 176.3 | 801 | 0.34 | 18 | 53 |
| *E. coli* lac I + trpB–trpA [22,23] | | 3 | 405 | 1.17 | 34 | 29 | | | | | |
| 7 Animal [11–17] | | 7 | 3786 | 0.33 | 90 | 269 | 248.9 | | | | |
| 5 Mammalian [13–17] | | 5 | 2379 | 0.35 | 65 | 184 | 169.8 | | | | |
| Chicken ovalbumin [12] | | | 1155 | 0.17 | 12 | 69 | 58.3 | 701 | 0.04 | 1 | 28 |
| Mouse $V\lambda_{II}$ Ig [13] | | | 495 | 0.10 | 2 | 21 | | | | | |
| Human beta globin [24] | | | | | | | | 136 | 0.11 | 1 | 9 |
| Human alpha globin [24] | | | | | | | | 146 | 0.40 | 4 | 10 |

*N* is the number of well defined genes in each sample. Expected CG or GC frequencies were calculated from base frequencies in each codon position. The overlap in SV40 contains one CG and six GC doublets; this accounts for the different values used here and in table 1. When the doublets are simply counted from one end of the total translated sequence to the other, the values given in this table are found. Because of SV40's overlap [3,28] and a few undetermined bases in $\phi$X174 [2], with these two viruses the sum of translated and untranslated bases does not exactly equal total bases. The *E. coli* sample is made up of only three partial sequences, in which not all codons were completely resolved [22,23], hence these genes cannot be considered well defined. The seven animal mRNAs have a CG/GC ratio slightly lower than that found for total genomes of other animal species. The 11 species analyzed [9] had CG/GC ratios for total DNA ranging from 0.40 (Coelacanth) to 0.85 (Thyone). The untranslated regions shown include both 5'- and 3'-ends

translated and untranslated regions. In SV40, however, the translated region contains 24-times as many GC doublets as CG doublets. The noncoding nucleotides of SV40 include nearly 3-times as many GC as CG doublets.

Codon frequencies provide a second index, intended to indicate the relative importance of nucleic acid and protein selections during evolution of a genome. Such an index is possible because of the choice available among the codons of arginine or serine, each of which has six codons. Arginine may be coded in mRNA by any member of the 'quartet' CGQ (Q is any one of the four mRNA bases, A, C, G or U) or the 'duet' AGX (X is either purine, A or G). A quartet is thus an ensemble of four codons for the same amino acid, each codon having the same initial doublet but a different third base. Likewise, a duet is an ensemble of two codons whose bases vary only in the third position (see fig.1). For arginine, our index is the ratio of CGQ to AGX codons. Assuming each codon equiprobable, the expected value for CGQ/AGX is 2.0 because of the number of codons in each kind of ensemble. Table 3 shows that arginine coding by the AGX duet is unusually frequent in SV40, since CGQ/AGX is only 0.08. Indeed, the SV40 AGX content of 4.10% of all codons is much higher than for any other sample.

A similar coding ratio is possible for serine. This amino acid is coded by the duet AGY (Y is either pyrimidine, C or U) and the quartet UCQ. Thus the AGY codons of serine have the same initial doublet as the AGX codons of arginine. (Leucine also possesses six codons, but the initial doublet varies in all cases from arginine or serine codons.) Table 3 reveals that in SV40 the UCQ/AGY ratio does not deviate nearly as much as does CGQ/AGX from the expected value calculated from individual base frequencies. Furthermore, in SV40 the serine coding ratio is close to that of animals while the arginine coding ratio is far removed. This contrast between values for the two

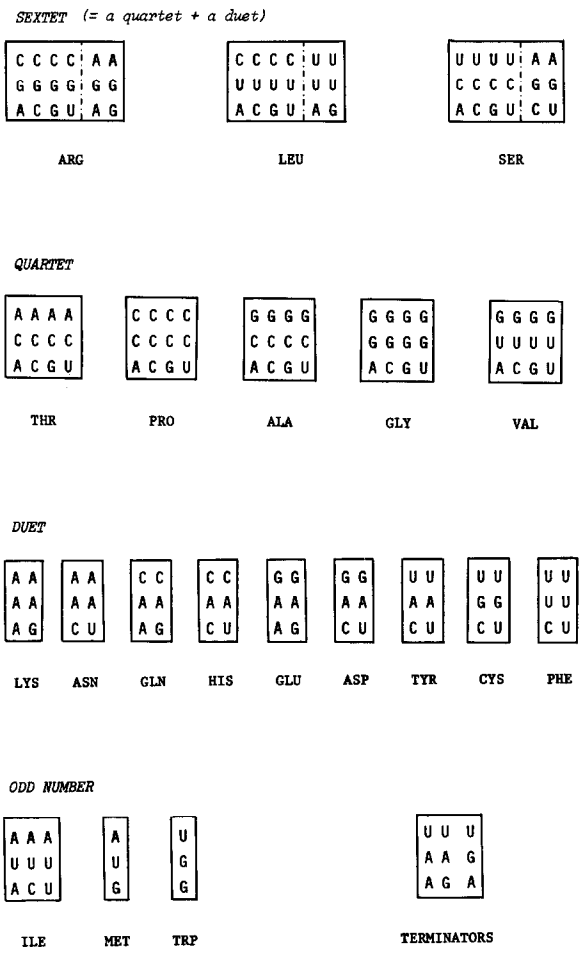


Fig.1. The genetic code. This convenient schema shows the 61 amino acid codons grouped into 20 ensembles of 1–6 members each. The international abbreviation for the amino acid coded appears below the ensemble. Note that the 3 amino acids with 'sextet' codons in fact have each a quartet and a duet of codons. Codons are read vertically. Restriction of amino acid coding by CG doublets in an mRNA sequence depends on codon positions occupied by the doublet. Each member of the Arg quartet (CGQ) carries a CG doublet in positions I–II. Position II–III CG doublets, however, only occur when G is used as third base in the quartets of Ser, Thr, Pro and Ala. CG doublets between two successive codons (positions III–I) can be formed only when the second codon has G in position I. Thus, absence of CG doublets between codons implies no Ala, Gly, Val, Glu or Asp coding by any codon following another ending in C. This and the position I–II effect from CGQ coding limit amino acid sequence possibilities. But CG absence in II–III prevents coding of no amino acid, hence imposes no constraint on the protein.

coding ratios would weigh against any neutral hypothesis seeking to explain codon frequencies in SV40, since no consistent selection of AGX and AGY duets relative to their cognate quartets is seen. The much greater variation of CGQ/AGX than UCQ/AGY among all these samples is a curious result of this study.

We now consider the question: Does the depressed CG doublet frequency in SV40 result from protein selection? The arginine content of SV40 total protein is 4.4%, nearly the same as that (4.5%) of the average protein for organisms [10]. The low CG doublet frequency is balanced by the high occurrence of AGX codons (but not generally of AG doublets! - - see below) and therefore cannot be due to selection of proteins low in arginine. Strong preference for AGX codons is peculiar to SV40, especially among viruses. Like their host, MS2 and $\phi$X174 both have higher mean frequencies per codon for CGQ members than for AGX members. That is, *Escherichia coli* and phages 'prefer' quartets to duets for coding arginine, as do several animal genes. In these cases, CGQ/AGX is elevated (see table 3).

Why is arginine nearly exclusively coded in SV40 by AGX instead of CGQ? The preference for AGX codons must be caused by contradictory protein and nucleic acid selection. A certain amount of arginine is evidently needed for proper functioning of SV40 proteins, yet this cannot be had by CGQ coding because of the CG doublet rarity. In fact, the CG doublet is more frequent in codon positions I–II than in positions II–III or III–I. The 4325 bases in the best-defined translated regions of the SV40 genome contain only nine CG doublets: five in positions I–II, four in III–I and none in II–III. Furthermore, the AG doublet, also, is more frequent in positions I–II than in II–III or III–I (observed to expected ratios in SV40 are 1.30, 0.99 and 0.95, respectively). In addition, table 3 shows 32% more AGX than AGY codons in SV40. This result is counter to the overall third base composition of SV40: Met and Trp codons aside, the remaining 1448 codons contain 46.9% A+G and 53.1% C+U in position III. For a given amount of CG and AG doublets, then, SV40 sequences tend to be arranged to maximize arginine coding! The reason for the AGX coding, therefore, apparently is that protein selection demands arginine while other constraints preclude CGQ coding.

No CG doublets in SV40 codon positions II–III

Table 3
Arginine and serine quartet-to-duet coding ratios

| | Arginine | | | | | Serine | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| mRNA sample | Total | %CGQ | %AGX | CGQ/AGX | expected | Total | %UCQ | %AGY | UCQ/AGY | expected |
| MS2 [1] | 7.04 | 5.82 | 1.22 | 4.8 | 2.24 | 9.10 | 6.85 | 2.25 | 3.0 | 2.10 |
| $\phi$X174 [2] | 5.49 | 5.00 | 0.49 | 10.1 | 1.97 | 6.78 | 5.98 | 0.80 | 7.5 | 2.34 |
| SV40 [3] | 4.43 | 0.33 | 4.10 | 0.08 | 1.33 | 6.49 | 3.44 | 3.04 | 1.1 | 1.45 |
| *E. coli* [22,23] | 5.26 | 5.26 | 0 | $\infty$ | | 6.02 | 5.26 | 0.75 | 7.0 | |
| 7 Animal [11–17] | 4.32 | 2.32 | 2.00 | 1.2 | 2.10 | 8.97 | 5.28 | 3.68 | 1.4 | 1.57 |
| 5 Mammalian [13–17] | 4.55 | 2.99 | 1.56 | 1.9 | 2.55 | 8.44 | 5.19 | 3.25 | 1.6 | 2.15 |
| *Ps. miliaris* H2B [11] | 4.76 | 4.76 | 0 | $\infty$ | | 7.14 | 3.57 | 3.57 | 1.0 | |
| Chicken ovalbumin [12] | 3.90 | 0.52 | 3.38 | 0.15 | | 9.87 | 5.97 | 3.90 | 1.5 | |
| Mouse immunoglobin [13] | 3.64 | 1.21 | 2.42 | 0.50 | | 11.52 | 7.27 | 4.24 | 1.7 | |
| Rat preproinsulin [14] | 5.05 | 5.05 | 0 | $\infty$ | | 3.03 | 2.02 | 1.01 | 2.0 | |
| Rat growth hormone [15] | 5.58 | 4.65 | 0.93 | 5.0 | | 7.44 | 3.72 | 3.72 | 1.0 | |
| Rabbit beta globin [16] | 2.05 | 0 | 2.05 | 0 | | 6.85 | 4.11 | 2.74 | 1.5 | |
| Human CS lactogen [17] | 5.36 | 3.57 | 1.79 | 1.99 | | 10.12 | 7.14 | 2.98 | 2.4 | |

Arginine or serine coding is expressed as % total codons (total number of codons = 1/3 number of bases in translated region of table 2). Arginine and serine each possess six codons, a 'quartet' and a 'duet' (see fig.1 and text). X = A or G; Y = C or U; Q = any of the four mRNA bases. Mean frequency per codon for the two kinds of ensembles is obtained by dividing the quartet or duet frequency given by 4 or 2, respectively. If each base had a frequency of 25%, the expected value for either CGQ/AGX or UCQ/AGY would be 2.0. The expected value shown is calculated from individual base frequencies in the sample. These are, for MS2: 23.6% A, 26.0% C, 25.6% G and 24.8% U [1]; for $\phi$X174: 23.7% A, 21.9% C, 23.2% G and 31.2% U [2]; for SV40: 31.5% A; 17.1% C, 23.3% G and 28.1% U [3]. The combined seven animal mRNAs have the following composition: 24.6% A, 26.4% C, 25.4% G and 23.6% U [11–17]. The five mammalian mRNAs have an overall base frequency distribution of: 22.2% A, 28.2% C, 26.4% G and 23.2% U [13–17]. The individual mRNAs are too small for comparing observed and expected codon frequencies

of course means zero frequencies for codons ACG, CCG, GCG and UCG [3]. II–III is the only codon position combination which imposes no constraint whatever on the protein by absence of CG doublets (see fig.1 legend). Hence, CG doublets can 'more easily' be eliminated from this combination and this is most likely the reason for the unequal distribution of CG doublets among the three position combinations. Although hardly used in the sea urchin H2B histone or chicken ovalbumin mRNA [11,12], the above four codons all appear several times in our combined mammalian sample [13–17]. Therefore, the absence of NCG codons in SV40 cannot be a consequence of a lack of appropriate tRNA for their translation in the host cell (for experimental indications see [18]).

Two explanations for the tiny overall CG doublet frequency in SV40 can be envisaged. The first is that the origin of the virus has determined the doublet frequencies. For example, SV40 could have originated relatively recently as a fragment of a larger genome rare in CG doublets in the part that gave rise to the virus. A remote origin of this nature must be excluded, however. In the absence of selection to keep CG doublets rare, mutations would gradually equilibrate CG and GC frequencies. A second possibility is direct selection against CG doublets during the evolution of SV40 itself, whatever its ancestor. In either case, we do not know why CG doublet rarity would be advantageous for SV40 or any other species. The great variation of CG/GC among species should ultimately make sense physicochemically. Unfortunately, experimental and theoretical estimates of the stability of CG and GC in nucleic acids do not agree [19–21]. Consequently, we cannot decide if an evolving polynucleotide rich in GC doublets but poor in CG doublets would be favored by economy of energy, stability or other properties.

It is too early to choose between hypotheses on the origin of SV40, yet some indications do exist. No other mRNA contains so low a CG doublet frequency as those of SV40. The combined five mammalian mRNA sequences show a CG/GC of 0.35, 8-times that of SV40's four genes. Mouse embryonic $V\lambda_{II}$ immunoglobin and rabbit beta globin mRNAs, however, do have rather low CG/GC ratios. Although the mouse messenger has the smallest CG/GC among genes so far sequenced in the animal kingdom (their range is 0.10–0.88), its value is more than double SV40's (see table 2). Thus by these data, SV40 contrasts strikingly with the coding nucleic acids of its host type cells. Conversely, MS2 and $\phi$X174 have CG/GC ratios grossly similar to that of *E. coli*, based on the scanty sequences published [22,23]. The viral messengers not included in table 1–3 can also be considered at this point. These are lambda cro, cII and 0; fd G3 and TMV coat [4–7]. The four bacteriophage mRNA make up a total of 311 codons having a CG/GC of 1.27 and a CGQ/AGX of 2.3, both indexes thus showing 'normal' values. The TMV partial coat mRNA of only 78 codons has a CG/GC of 1.09 and a CGQ/AGX of 0.17. Of all coding and noncoding sequences determined, the only one with a CG/GC comparable to SV40 is the 3′ untranslated region following the chicken ovalbumin mRNA [12] (see table 2).

Although CG doublets are rarer in SV40 genes than in any other genes sequenced, the same is not true of untranslated regions. Human beta globin 5′- and 3′-ends have a CG/GC ratio of 0.11 in their 131 noncoding nucleotides, much lower than the 0.34 for SV40 untranslated bases. Incidentally, the beta globin pattern is not followed by alpha globin, either in translated or untranslated regions. Although its sequence is not completed, the alpha mRNA has more CG doublets in both regions than does beta; furthermore, the CGQ quartet is not avoided in alpha as it is in beta [24] (see table 2 for noncoding zones). Curiously, human papovavirus BKV may have a lower CG/GC in noncoding zones than does SV40. BKV has very few CG doublets in the untranslated region believed homologous to nucleotides 5101–5228 of the SV40 sequence. In this interval, which includes the DNA replication origin of both sequences, BKV has 2 CG doublets and 11 GC doublets among 155 bases, while SV40 has 4 CG and 14 GC doublets among 127 nucleotides. This part of the sequence, therefore, shows a CG/GC ratio of 0.18 for BKV and 0.29 for SV40 [3,25].

A synthesis of gene comparisons will now be attempted. However, for more complete mRNA characterization, two new indexes are added. The first is the G-ending codon ratio NCG/NMG, where N is any base and M any except C. The other index is %C+%G of codon position III. Values for these indexes, as well as CG/GC and CGQ/AGX, appear in table 4. Also shown, but not used as an mRNA index

since it is determined by amino acid choice in protein, is %C+%G in combined codon positions I and II. Each gene appears individually and each ratio is expressed in absolute frequencies. Thus, the first line of the table reveals that the big T antigen messenger of 626 codons tops the list in two respects. The NCG/NMG

Table 4
Summary of gene indexes for sequenced mRNA

| Gene | NCG/NMG | CG/GC | CGQ/AGX | %C+%G in III | %C+%G in I+II | Ranks | Sum |
|---|---|---|---|---|---|---|---|
| SV40 T (4489–2611) | 0/134 | 2/82 | 2/22 | 34.3 | 39.6 | 1, 1, 4, 4 | 10 |
| SV40 VP1 (1426–2509) | 0/79 | 2/49 | 1/12 | 36.9 | 46.1 | 2, 2, 3, 6 | 13 |
| SV40 VP2 (483–1536) | 0/59 | 5/65 | 2/20 | 29.1 | 50.4 | 3, 4, 5, 1 | 13 |
| SV40 t (5079–4560) | 0/39 | 1/23 | 0/8 | 37.0 | 39.6 | 5, 3, 1, 7 | 16 |
| SV40 VP3 (837–1536) | 0/52 | 3/31 | 2/19 | 32.2 | 45.9 | 4, 6, 6, 2 | 18 |
| Mouse $V\lambda_{II}$ Ig (165) | 0/20 | 2/21 | 2/4 | 40.6 | 51.2 | 7, 5, 9, 10 | 31 |
| Chick ovalbumin (385) | 1/79 | 12/69 | 2/13 | 46.8 | 44.2 | 8, 8, 7, 18 | 41 |
| Rabbit $\beta$ globin (146) | 1/54 | 4/33 | 0/3 | 65.1 | 49.0 | 9, 7, 2, 25 | 43 |
| $\phi$X174 H (326) | 8/52 | 32/78 | 6/1 | 34.0 | 50.6 | 14, 10, 17.5, 3 | 44.5 |
| Lambda 0 (98) | 1/20 | 12/19 | 4/4 | 44.9 | 43.4 | 10, 13, 11, 17 | 51 |
| G4 K (55) | 2/11 | 7/9 | 2/0 | 38.2 | 41.8 | 15, 16, 19, 9 | 59 |
| $\phi$X174 A (511) | 22/108 | 79/104 | 28/7 | 44.8 | 44.7 | 16, 15, 15, 16 | 62 |
| Lambda cII (96) | 7/25 | 20/24 | 7/2 | 41.7 | 49.5 | 17, 17, 14, 14 | 62 |
| Rat gr. horm. (215) | 4/69 | 23/70 | 10/2 | 73.5 | 49.3 | 11, 9, 16, 27 | 63 |
| Human CSL (168) | 7/52 | 21/37 | 6/3 | 79.8 | 42.3 | 13, 12, 12, 29 | 66 |
| Lambda cro (65) | 3/8 | 11/17 | 3/0 | 41.5 | 44.6 | 20, 14, 21, 13 | 68 |
| Rat preproins. (99) | 2/31 | 11/23 | 5/0 | 69.7 | 55.6 | 12, 11, 25.5, 26 | 74.5 |
| TMV coat (78) | 8/9 | 13/11 | 1/6 | 41.0 | 46.2 | 29, 27, 8, 11 | 75 |
| Sea urchin H2B (84) | 0/23 | 14/16 | 4/0 | 75.0 | 45.8 | 6, 18, 23.5, 28 | 75.5 |
| $\phi$X174 K (55) | 3/10 | 10/10 | 3/0 | 41.8 | 40.0 | 19, 22, 21, 15 | 77 |
| fd G3 (72) | 6/10 | 17/19 | 0/0 | 55.6 | 49.3 | 27, 20, 10, 21 | 78 |
| $\phi$X174 G (174) | 9/21 | 24/23 | 5/0 | 36.2 | 46.3 | 26, 23, 25.5, 5 | 79.5 |
| $\phi$X714 F (422) | 23/55 | 66/75 | 25/0 | 37.9 | 49.1 | 25, 19, 29, 8 | 81 |
| MS2 coat (129) | 5/17 | 27/30 | 4/0 | 51.2 | 49.2 | 18, 21, 23.5, 19 | 81.5 |
| MS2 A (392) | 29/77 | 90/72 | 22/7 | 56.3 | 50.3 | 21, 29, 13, 22 | 85 |
| $\phi$X174 D (151) | 7/18 | 36/32 | 11/0 | 41.1 | 50.3 | 23, 24, 28, 12 | 87 |
| MS2 replicase (544) | 38/94 | 128/105 | 36/6 | 54.3 | 50.2 | 24, 28, 17.5, 20 | 89.5 |
| *E. coli* lac I (61) | 8/13 | 17/15 | 3/0 | 62.0 | 50.0 | 28, 25, 21, 23 | 97 |
| $\phi$X174 E (90) | 11/29 | 23/20 | 6/0 | 63.3 | 40.6 | 22, 26, 27, 24 | 99 |

Each gene is ranked by four indexes. These are: the G-ending codon ratio, NCG/NMG, where N is any base and M is any base except C; the CG/GC doublet ratio; the arginine coding ratio, CGQ/AGX, where CGQ is the quartet that codes arginine and AGX is its duet (see text and fig.1); and %C+%G in codon position III. The nucleotide interval or number of codons in the sequence appears in parenthesis after the gene designation. The first three indexes are expressed in absolute frequencies. For example, the CGQ/AGX for SV40 T is 2/22; this means that in the nucleotide interval shown, 2 CGQ and 22 AGX codons occur. Not all sequences are definitely known to be translated; in particular, mouse $V\lambda_{II}$ immunoglobin is uncertain [13]. $\phi$X174 gene J is excluded since it contains only 37 codons, but the frameshift genes E and K appear [2,29]. SV40 T is not fully resolved, the 626 most certain codons are taken. SV40 VP3 is formed by the last (3′) 233 codons of VP2. The main reason for the difference in %C+%G of codon positions I+II, also shown in the table, between VP2 and VP3 is that, of the first 93 VP2 triplets, 29 code Ala (of these, 24 are GCU codons). The C+G value in positions I+II is 64% for the initial 93 codons of VP2 [3,28,30]. %C+%G of I+II is a protein derived characteristic; it is not ranked and does not count as an mRNA index. Ranking is somewhat arbitrary in a few cases. For instance, bacteriophage fd G3 mRNA contains no Arg codons [6]; its CGQ/AGX is therefore 0/0, which is ranked immediately before the 4/4 of phage lambda [7]. To save space references are omitted since they have all already been given except for MS2 coat and A mRNA [32,33]

of 0/134 is the most extreme of the 29 genes; this means that T's mRNA has no G-ending codons with C as middle base but does have 134 such codons with A, G or U as middle base. The second ratio, CG/GC, also places this gene first since its 1878 nucleotides contain only 2 CG doublets but 82 GC doublets. The third index ranks T fourth because little t, rabbit beta globin and SV40 VP1 mRNAs have smaller CGQ/AGX ratios. Next, T is ranked fourth also for lowness in %C+%G of third bases. Finally, note that the 'protein index' %C+%G in I+II assigns the lowest values to T and t. This is a small indication that the early genes T and t have exceptional proteins as well as nucleic acid sequences.

These indexes are fairly independent. Omitting SV40, which we can consider a test case, the only strong correlation among table 4 ranks is between the first two ratios, NCG/NMG and CG/GC ($r = 0.81$). The third ratio, CGQ/AGX, correlates weakly with NCG/NMG ($r = 0.34$) and CG/GC ($r = 0.41$). All three ratios correlate negatively, although weakly, with %C+%G of third bases ($r = -0.28$, $-0.03$ and $-0.15$, respectively). Correlation of each of these four indexes with %C+%G in I+II is very weak ($r = +0.10$, $-0.02$, $-0.03$ and $-0.09$, respectively). Consequently, it is interesting that the four 'nucleic acid indexes' give such similar rankings for all SV40 genes.

A rough overall comparison of gene similarity is obtained by adding the ranks for these indexes. The resulting sums give a simplified and, possibly, a biased picture of differences between genes. This is, however, a good way of showing variation among genes. By the rank sums (last column of table 4), SV40 genes are clearly at one end of a spectrum and are separated by appreciable distances from other genes. (This can be seen by constructing a frequency diagram against rank sum.) Genes appear in table 4 in the order of these sums. A further qualification on the above question of SV40's origin is now possible: A recent origin as fragment of another genome seems unlikely in the light of several kinds of differences between SV40 genes and the other 24 genes. In some cases, especially with NCG/NMG, no more extreme example can hardly be expected to be found among future gene or genome sequences. The implication is that the peculiar characteristics of SV40 must have been selected during evolution of that virus itself.

Finally, the use of degenerate bases in SV40 deserves special mention. One of the surprises to emerge from sequencing mammalian mRNA is their high content of C and G in codon position III. The five mammalian genes have much greater values than SV40 for %C+%G of third bases. (Mouse $V\lambda_{II}$ Ig mRNA, however, shows only 40.6% [13].) The range among SV40 genes is only 29.1–37.0 for %C+%G in III. Also, the third base frequency order is C<G<A<U for all its genes. The greatest relative difference between SV40 and mammalian mRNA is with bases A and C (see table 5).

Since codon and anticodon amounts are believed to be coordinated in the cell [18], the question of optimum translation of SV40 codons, particularly the A-ending codons, is raised. Not only is %A in position III elevated, it is nearly constant among SV40 genes (table 5). Judging from third base compositions, the chicken ovalbumin system would seem better for translation of SV40 mRNA than mammalian cells! It would be surprising if the same isoacceptor tRNA population could efficiently decode mammalian and SV40 messengers. This is so since each A-ending duet codon (see fig.1) has a tRNA specific for that codon alone. These tRNA species all contain U* (modified U) in the anticodon for recognizing A in position III of the codon. It is therefore reasonable to expect that the cell has an optimum balance between the number of A-ending duet codons and such tRNA with U* in the anticodon. These same tRNA must decode SV40 mRNA since no tRNA genes exist in its genome. Recent observations indicate that tRNA specific for A-ending duet codons are generally rarer than those for G-ending duet codons in eukaryotes (G. Chavancy, A. Chevallier, A. Fournier and J.-P. Garel, in preparation).

That the mammalian cell is deficient in tRNA for translating the frequent A-ending codons of SV40 is implied in the following comparison: In SV40, the A-ending duet codons of Arg, Leu, Lys, Gln and Glu make up 16.5% of all codons [3], but in our mammalian sample they represent only 5.4%, one-third as much. The case of tRNAs for quartet codons is less clear. I is found in the anticodons for such codons (except those of Leu and Gly [26]), and I is ordinarily assumed to have a decoding efficiency of C>U>A. In SV40, %A of position III is about as high in quartets as in all codons, while in mammals it is about as low in quartets as in all codons (table 5). A-ending mem-

Table 5
Use of degenerate codon bases

| Sample | % in position III | | | | % in III of quartets | | | |
|---|---|---|---|---|---|---|---|---|
| | C | G | A | U | C | G | A | U |
| MS2 (3 genes) | 30.2 | 24.4 | 19.4 | 25.9 | 27.5 | 22.3 | 19.9 | 30.3 |
| $\phi$X174 (6 genes) | 18.9 | 20.3 | 13.6 | 47.2 | 20.3 | 13.2 | 9.8 | 56.7 |
| SV40 (4 genes) | 13.4 | 20.6 | 28.5 | 37.5 | 13.8 | 13.3 | 27.7 | 45.2 |
| VP1 | 15.0 | 21.9 | 28.0 | 35.2 | 13.3 | 15.8 | 29.1 | 41.8 |
| VP2 | 12.3 | 16.8 | 28.5 | 42.5 | 14.5 | 9.6 | 24.1 | 51.8 |
| T | 12.9 | 21.4 | 28.3 | 37.4 | 15.0 | 14.5 | 28.5 | 42.0 |
| t | 14.5 | 22.5 | 30.6 | 32.4 | 9.3 | 13.0 | 31.5 | 46.3 |
| 7 Animal mRNA | 33.5 | 27.2 | 16.2 | 23.1 | 33.6 | 23.2 | 17.2 | 25.9 |
| 5 Mammalian mRNA | 35.7 | 30.3 | 12.0 | 22.1 | 34.9 | 28.1 | 12.5 | 24.5 |
| *Ps. miliaris* H2B | 47.6 | 27.4 | 8.3 | 16.7 | 52.4 | 14.3 | 11.9 | 21.4 |
| Chick ovalbumin | 26.0 | 20.8 | 26.5 | 26.8 | 25.6 | 14.0 | 29.9 | 30.5 |
| Mouse $V\lambda_{II}$ Ig | 28.5 | 12.1 | 24.8 | 34.5 | 27.7 | 6.4 | 28.7 | 37.2 |
| Rat preproinsulin | 36.4 | 33.3 | 10.1 | 20.2 | 30.2 | 34.0 | 7.5 | 28.3 |
| Rat growth hormone | 39.5 | 34.0 | 8.4 | 18.1 | 44.0 | 29.0 | 7.0 | 20.0 |
| Rabbit beta globin | 27.4 | 37.7 | 6.2 | 28.8 | 27.6 | 39.5 | 2.6 | 30.3 |
| Human CS lactogen | 44.6 | 35.1 | 10.1 | 10.1 | 43.5 | 39.1 | 13.0 | 4.3 |

Samples are the same as in table 1. The base composition of codon position III is shown for all codons (left side of table) and for quartet codons only (right side, see fig.1 for identity of the eight quartets). Bases appear in the frequency order found for all SV40 genes, but for no other gene sequenced, C<G<A<U. Note that in SV40 %C is about the same in all codons or quartets only while G is lower in quartets. This could be a consequence of general pressure for eliminating CG doublets from the translated part of the genome. Eliminating G from position III of only the codons of Ser, Thr, Pro and Ala, avoids CG doublets in positions II–III. C as third base, however, needs to be eliminated from codons of 15 different amino acids to avoid CG doublets in positions III–I (i.e., between successive codons). C can be replaced in all 15 cases without changing the protein, but replacing G in codon position I implies amino acid substitution. Hence the only strategy for lowering CG doublet frequency in III–I without affecting the protein is to lower %C in position III. This elimination is among all categories of codons while that for G is among quartet codons only. Thus we find %C similar in quartets or all codons while %G is considerably lower in quartets. And C is the rarest third base in SV40 because its elimination involves codons of 75% of all the amino acids

bers of quartet and Ile ensembles account for 12.0% of all SV40 codons, but only 6.6% of the mammalian sample. Direct measurement of the tRNA population in SV40 infected cells (both permissive and non-permissive) should help to clarify this situation. In waiting for such experiments, I suggest that if a large quantity of SV40 mRNA were to bind to ribosomes the abundant A-ending SV40 codons would tie up the corresponding tRNA of the cell, thus inhibiting synthesis of cell proteins. For example, suppose SV40 mRNA is added to a system synthesizing rabbit beta globin until the total number of codons from each species is equal. The resulting slowing of amino acid incorporation into beta globin can be expected to be much more than a factor of two because the appropriate tRNA would collide with A-ending codons of SV40 over 4-times as often as with those of the beta globin messenger. Although the total number of codons is equal for the two kinds of mRNA, A-ending codons represent 28.5% of all SV40 codons but merely 6.2%

of beta globin codons. In other words, 4.6-times as many A-ending codons of SV40 would be present as of the rabbit messenger.

The relatively low amount of functional SV40 mRNA (~5% [27]) found compared to total cell mRNA may reflect this tRNA shortage. Perhaps the cell could not remain living with a higher content of functioning SV40 mRNA, especially if the idea that tRNA is limiting has any validity. How SV40 mRNA is held to only ~5% of cell mRNA is a mystery since an enormous quantity of SV40 DNA is present. Some kind of feedback between translation inhibition and RNA polymerase II exploitation may be involved.

Finally, a strange phenomenon remains regarding doublet distribution among codon position combinations. Like SV40, 20 of the 24 other mRNA show fewer CG doublets in positions II–III (protein constraint free) than in I–II (arginine coding). The four exceptions are fd G3, $\phi$X174 K, TMV coat and rabbit beta globin. But of the 20, only four (G4 K, *E. coli* lac I, chick ovalbumin and human CSL) also have more AG doublets in I–II than in II–III. Maximization of arginine coding by concentration of both CG and AG doublets in positions I–II is, therefore, rather rare, but does occur in other species than SV40 (data not shown).

This short review has raised more questions than it has answered, but some of the questions should stimulate future research, especially since partial answers have been attempted: What is the nature of the selection that has led to the total lack of NCG codons in SV40? Why are A-ending codons so frequent in SV40 and so rare in its host? What would happen in a primate cell having an iso-tRNA population optimum for SV40 messengers? Would much more viral mRNA be transcribed and translated? Why do mammalian genes use so much C and G in codon position III and SV40 genes so little? In sum, why is SV40 so different from its host by the above four indexes? Lastly, is it possible that SV40 came from birds? The 637 untranslated 3′ bases of the chicken ovalbumin gene contain no CG doublets [12]. The best candidate to date for ancestor of SV40 is therefore a noncoding region in the chicken genome!

## 3. Conclusion

Four indexes have been applied to mRNA sequences to compare genes and genomes. These are the doublet ratio CG/GC, the arginine coding ratio CGQ/AGX, the G-ending codon ratio NCG/NMG, and %C+%G in codon position III. Of the 29 mRNA studied, those of SV40 show the lowest values for each index.

The small but varying amount of CG doublets in the three possible codon position combinations may signal a kind of mRNA molecular selection. This is suggested by the much lower frequency of CG doublets in the SV40 coding region. Such a difference between coding and noncoding regions could, of course, simply mean selection of a sequence of codons that minimized arginine coding. This would be an untenable conclusion in the present case, however. In SV40 genes, CG and AG doublets are both concentrated in positions I–II, thus increasing arginine coding. Furthermore, most AG doublets in positions I–II are followed by a purine third base, that is they code arginine rather than serine. The overall rarity of CG doublets in SV40 mRNA must therefore be due to selection against the CG doublet itself and not against its ability to code arginine. The much stronger discrimination against CG doublets in mRNA of this molecular parasite of the mammalian cell than in the mRNA native to that cell is puzzling. Many species have lower than expected CG/GC ratios and this suggests an importance throughout evolution in maintaining low CG doublet frequency. SV40 is so extreme compared to its host type cell in this respect that understanding should come from working with it.

The SV40 genome seems much more efficient in replication than in transcription since a relatively enormous amount of its DNA is replicated but only a small quantity is transcribed. This may be related to the apparent rareness in the mammalian cell of tRNA for decoding the frequent A-ending codons in SV40 mRNA, but experiments must be done to verify and clarify the picture.

Both mammalian and SV40 nucleic acids exhibit evidence of strong selection among degenerate bases. But this protein-independent bias is opposite in sense in the two cases. The bases with the lowest degenerate frequency in SV40, C and G, most often have the highest degenerate frequency in mammalian mRNA. The reason for the great variation of C+G content in codon position III among species is unclear. The consistently low %C+%G of SV40 third bases, however, seems to reflect general selection against the CG

doublet in genes of this virus.

The indexes presented constitute a new tool for investigating possible origins of viruses or selective constraints (co-evolution) between host and parasite. In general, my results imply a strong contribution from nucleic acid selection (as independent of protein selection) in the evolution of SV40. I believe this analysis tends to favor a polyphyletic origin for viruses. For the time being, the question of the origin of SV40 must be left open; however, possibilities are constrained somewhat by the above analysis. Finally, it seems to me that SV40 must either represent a bizarre case of co-evolution or else be a recent invader of the mammalian cell.

## Acknowledgements

I thank C. Biémont, G. Chavancy, J. David, J.-P. Garel, M. Gouy and J.-J. Madjar for help of various kinds, including valuable criticism and suggestions.

## References

[1] Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G. and Ysebaert, M. (1976) Nature 260, 500–507.

[2] Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison III, C. A., Slocombe, P. M. and Smith, M. (1977) Nature 265, 687–695.

[3] Reddy, V. B., Thimmappaya, B., Dhar, R., Subramanian, K. N., Zain, B. S., Pan, J., Ghosh, P. K., Celma, M. L. and Weissman, S. M. (1978) Science 200, 494–502.

[4] Guilley, H., Jonard, G., Richards, K. E. and Hirth, L. (1975) Eur. J. Biochem. 54, 145–153.

[5] Roberts, T. M., Shimatake, H., Brady, C. and Rosenberg, M. (1977) Nature 270, 274–275.

[6] Sugimoto, K., Sugusaki, H., Okamoto, T. and Takanami, M. (1977) J. Mol. Biol. 110, 487–507.

[7] Schwarz, E., Scherer, G., Hobom, G. and Kössel, H. (1978) Nature 272, 410–414.

[8] Grantham, R. (1972) Nature New Biol. 237, 265–266.

[9] Russell, G. J. and Subak-Sharpe, J. H. (1977) Nature 266, 533–535.

[10] Dayhoff, M. O. (1976) Atlas of Protein Sequence and Structure, vol. 5, suppl. 2, p. 301, Georgetown University Medical Center, Natl. Biomed. Res. Found., Washington, DC.

[11] Birnstiel, M. L., Schaffner, W. and Smith, H. O. (1977) Nature 266, 603–607.

[12] McReynolds, L., O'Malley, B. W., Nisbet, A. D., Fothergill, J. E., Givol, D., Fields, S., Robertson, M. and Brownlee, G. G. (1978) Nature 273, 723–728.

[13] Tonegawa, S., Maxam, A. M., Tizard, R., Bernard, O. and Gilbert, W. (1978) Proc. Natl. Acad. Sci. USA 75, 1485–1489.

[14] Ullrich, A., Shine, J., Chirgwin, J., Pictet, R., Tischer, E., Rutter, W. J. and Goodman, H. M. (1977) Science 196, 1313–1319.

[15] Seeburg, P. H., Shine, J., Martial, J. A., Baxter, J. D. and Goodman, H. M. (1977) Nature 270, 486–494.

[16] Efstradiadis, A., Kafatos, F. C. and Maniatis, T. (1977) Cell 10, 571–585.

[17] Shine, J., Seeburg, P. H., Martial, J. A., Baxter, J. D. and Goodman, H. M. (1977) Nature 270, 494–499.

[18] Garel, J.-P. (1976) Nature 260, 805–806.

[19] Ninio, J. (1971) J. Mol. Biol. 56, 63–82 (see P. Claverie appendix, pp. 75–82).

[20] Tinoco, I., jr, Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M. and Gralla, J. (1973) Nature New Biol. 246, 40–41.

[21] Borer, P. N., Dengler, B., Tinoco, I., jr and Uhlenbeck, O. C. (1974) J. Mol. Biol. 86, 843–853.

[22] Platt, T. and Yanofsky, C. (1975) Proc. Natl. Acad. Sci. USA 72, 2399–2403.

[23] Steege, D. A. (1977) Proc. Natl. Acad. Sci. USA 74, 4163–4167.

[24] Baralle, F. E. (1977) Cell 12, 1085–1095.

[25] Dhar, R., Lai, C.-J. and Khoury, G. (1978) Cell 13, 345–358.

[26] Ninio, J. (1973) Prog. Nucl. Acid Res. Mol. Biol. 13, 301–337.

[27] Eckhart, W. (1974) Ann. Rev. Gen. 8, 302.

[28] Contreras, R., Rogiers, R., Van de Voorde, A. and Fiers, W. (1977) Cell 12, 529–538.

[29] Shaw, D. C., Walker, J. E., Northrop, F. D., Barrell, B. G., Godson, G. N. and Fiddes, J. C. (1978) Nature 272, 510–515.

[30] Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G. and Ysebaert, M. (1978) Nature 273, 113–120.

[31] Volckaert, G., Van de Voorde, A. and Fiers, W. (1978) Proc. Natl. Acad. Sci. USA 75, 2160–2164.

[32] Min Jou, W., Haegeman, G., Ysebaert, M. and Fiers, W. (1972) Nature 237, 82–88.

[33] Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Merregaert, J., Min Jou, W., Raeymakers, A., Volkaert, G., Ysebaert, M., Van de Kerckhove, J., Nolf, F. and Van Montagu, M. (1975) Nature 256, 273–278.