# NUCLEIC ACID SEQUENCE SIMILARITIES: 'POLY(A) TENDENCY'

Richard GRANTHAM

*Equipe Evolution Moléculaire, Laboratoire de Biométrie, Université Lyon I, 69622 Villeurbanne Cedex, France*

Received 14 August 1980
Revised version received 30 September 1980

## 1. Introduction

The origin and evolution of viruses is coming under stronger attack, thanks to the increasing availability of gene and genome sequences. However, the methodology is still narrow, being mainly based on alignment of sequences. In this paper I continue the effort to develop indexes for characterizing nucleic acid sequences and establishing relatedness among them [1].

As previously shown, the DNA sequences of papova virus SV40 and the untranslated zone following the chicken ovalbumin gene are similar in having very low frequencies for the dinucleotide CG [1]. An additional likeness is here described in 'poly(A) tendency'. A sequence with poly(A) tendency has an elevated frequency for purely adenine (A) containing oligonucleotides of 4 or more bases in length. Two sample sequences high in runs of A are compared to all published mRNA sequences [2] and to a large number of untranslated sequences.

AAAA is the most frequent of the 256 possible tetranucleotides in the 637 untranslated bases 3' to the chicken ovalbumin codons [3] (637ov3') and in 4 papova virus genes, as revealed below. Although I find poly(A) tendency throughout the SV40 and BKV genomes [4–7], the VP1 gene is most like 637ov3' in this respect. The maximum tendency in SV40 (the two papova genomes are very similar in poly(A) tendency, see below) is reached in the first half of VP1.

Table 1
Observed and theoretical runs of adenine

| Sample | | Number of A in | | | | | | | | $A_n$ ($n > 3$) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $A_6$ | $A_7$ | $A_8$ | %Total A | %Total bases |
| 637ov3' (637 | O | 77 | 58 | 27 | 16 | 30 | 0 | 0 | 8 | 25.0 | 8.5 |
| bases, 216 A) | T | 94.7 | 64.1 | 32.4 | 14.6 | 6.1 | 2.5 | 1.0 | 0.4 | 11.4 | 3.9 |
| 636 vp1 (636 | O | 69 | 48 | 33 | 28 | 15 | 18 | 0 | 0 | 28.9 | 9.6 |
| bases, 211 A) | T | 94.6 | 62.7 | 31.0 | 13.6 | 5.6 | 2.2 | 0.8 | 0.3 | 10.7 | 3.5 |
| SV40 VP1 (1083 | O | 139 | 78 | 42 | 40 | 25 | 18 | 0 | 0 | 24.3 | 7.7 |
| bases, 342 A) | T | 160.5 | 101.2 | 47.8 | 20.0 | 7.8 | 2.9 | 0.8 | 0.4 | 9.3 | 3.0 |
| *Sac.* CC1 (324 | O | 38 | 32 | 15 | 8 | 10 | 6 | 0 | 0 | 22.0 | 7.4 |
| bases, 109 A) | T | 48.4 | 32.4 | 16.2 | 7.2 | 3.0 | 1.0 | 0.4 | 0.2 | 10.8 | 3.6 |

The number observed (O) of adenines (A) in runs of all lengths is shown. Theoretical values (T) are calculated from the distribution theory for runs of given length [8]. 637ov3' is the untranslated 637 base sequence 3' to ovalbumin codons in the chicken [3]. 636vp1 contains the first 636 bases of SV40 VP1; the whole gene sequence [4,5] is also analyzed for comparison. *Sac.* CCl, the *Saccaromyces* cytochrome *c* mRNA [11] has the strongest poly(A) tendency of any messenger outside papova viruses. For 637ov3', %$A_n$ ($n > 3$) = 100(16 + 30 + 8)/637, or 8,5%, and % total A in $A_n$ = 100(54/216) = 25.0

I therefore take the first 212 codons as a sample (636vp1) of about the same length as 637ov3′ for comparison. For runs of 4 or more A ($A_n$ with $n > 3$), the observed frequency is over double the theoretical frequency [8] with either sample, as seen in table 1.

I have analyzed the 119 published mRNA sequences [2,9] and find poly(A) tendency in certain other double stranded DNA viruses and even in some single-stranded DNA bacteriophages. The effect also exists in single-stranded RNA plant viruses. A few bacterial and animal genes as well exhibit elevated contents of adenine runs. Curiously, the same phenomenon is not seen with all bases. Table 2 shows all mRNA

Table 2
Messenger RNA sequences with high poly A tendency

| Sequence | No. bases | Overall composition (%) | | | | Most frequent quadruplet (no.) | | For runs with $n > 3$ | | | | [Ref.] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A | C | G | U | | | %$A_n$ | %$C_n$ | %$G_n$ | %$U_n$ | |
| 636vp1 | 636 | 33.2 | 18.7 | 23.3 | 24.8 | AAAA | (22) | 9.6 | 1.3 | 1.9 | 3.8 | [4] |
| 637ov3′ | 637 | 33.8 | 19.5 | 17.0 | 29.7 | AAAA | (21) | 8.5 | 0 | 1.4 | 1.9 | [3] |
| Chick ovalb. | 1155 | 29.4 | 21.6 | 23.4 | 25.6 | AGAG | (16) | 1.6 | 0.7 | 0 | 0.7 | [3] |
| TMV PSV 30K | 315 | 34.9 | 12.1 | 26.0 | 27.0 | AAAA, AGAA | (7) | 6.0 | 0 | 0 | 2.5 | [30] |
| TYMV coat | 564 | 22.9 | 39.4 | 16.5 | 21.3 | CACC | (15) | 0 | 5.1 | 0 | 0 | [31] |
| $\phi$X174 B | 357 | 31.4 | 23.2 | 22.4 | 23.0 | AAAA | (8) | 6.4 | 0 | 0 | 2.2 | [32] |
| G | 522 | 19.9 | 22.6 | 20.3 | 37.2 | UGGU | (13) | 0 | 0.8 | 0 | 6.3 | [32] |
| G4 H | 1008 | 28.9 | 25.5 | 21.7 | 23.9 | CAAA | (19) | 5.4 | 0.4 | 0 | 0.8 | [33] |
| fd 1 | 1041 | 28.9 | 16.8 | 19.1 | 35.2 | AAAA | (18) | 5.2 | 0.4 | 0 | 2.5 | [34] |
| 2 | 1227 | 25.9 | 20.7 | 18.2 | 35.2 | AUUU, UUUG, UUUU | (18) | 3.8 | 0.7 | 0.4 | 4.6 | [34] |
| 3 | 1269 | 22.9 | 21.3 | 24.6 | 31.2 | GGUG | (24) | 4.1 | 0.6 | 0.4 | 1.6 | [34] |
| 4 | 1275 | 25.5 | 18.6 | 19.6 | 36.3 | UGUU | (23) | 2.4 | 0.3 | 0 | 4.9 | [34] |
| 6 | 333 | 19.2 | 19.5 | 17.7 | 43.5 | UAUU | (11) | 3.3 | 0 | 0 | 6.6 | [34] |
| 10 | 330 | 23.3 | 21.2 | 17.6 | 37.9 | UUCC | (7) | 3.6 | 1.5 | 0 | 6.4 | [34] |
| M13 1 | 645 | 25.9 | 17.4 | 21.9 | 34.9 | AUUA, UUCU, UUGU | (10) | 4.5 | 0 | 0 | 2.8 | [35] |
| 3 | 1269 | 23.0 | 21.4 | 24.5 | 31.0 | GGUG | (22) | 4.0 | 1.0 | 0.8 | 1.9 | [35] |
| 6 | 333 | 18.9 | 19.2 | 17.7 | 44.1 | UUAU, UUCU | (8) | 3.3 | 0 | 0 | 6.6 | [35] |
| λ cI | 708 | 29.4 | 20.5 | 24.9 | 25.3 | AAAA | (14) | 4.1 | 0 | 2.4 | 1.6 | [36] |
| 0 | 894 | 32.2 | 23.5 | 24.7 | 19.6 | AAAA | (19) | 7.2 | 0.4 | 2.1 | 0.4 | [37] |
| P1CmCAT | 654 | 26.5 | 22.0 | 22.9 | 28.6 | AAAA | (11) | 2.4 | 0.8 | 0.6 | 5.8 | [38] |
| SV40 T | 1878 | 32.9 | 15.5 | 22.4 | 29.2 | AAAA | (45) | 6.6 | 0.9 | 1.0 | 3.9 | [4] |
| t | 519 | 32.6 | 14.5 | 24.3 | 28.7 | AAAA | (13) | 7.1 | 0 | 1.0 | 0.8 | [4] |
| vp1 | 1083 | 31.6 | 19.0 | 24.0 | 25.4 | AAAA | (29) | 7.7 | 1.1 | 1.8 | 3.8 | [4] |
| vp3 | 699 | 31.5 | 18.9 | 22.3 | 27.3 | CAAA | (12) | 4.6 | 1.1 | 1.1 | 2.7 | [4] |
| BKV t | 513 | 30.6 | 18.3 | 24.0 | 27.1 | AAAA | (10) | 5.5 | 0 | 1.8 | 3.3 | [6] |
| *E. coli* rL11 | 423 | 26.5 | 26.0 | 26.7 | 20.8 | AAAA | (8) | 4.7 | 1.0 | 0 | 0 | [39] |
| rL12 | 360 | 30.3 | 19.2 | 27.2 | 23.3 | GCUG | (15) | 4.7 | 0 | 0 | 0 | [39] |
| Yeast CC1 | 324 | 33.6 | 19.1 | 23.8 | 23.5 | AAAA | (9) | 7.4 | 0 | 0 | 1.2 | [11] |
| *B. mori* fib. | 498 | 31.9 | 17.9 | 25.1 | 25.1 | AAAA | (15) | 6.6 | 0 | 0.8 | 1.6 | [40] |
| Mus Ig K2 | 351 | 27.9 | 23.6 | 22.5 | 25.9 | UCAA, UUUU | (6) | 2.6 | 0 | 1.1 | 5.1 | [41] |
| Rab. α globin | 423 | 20.6 | 35.7 | 26.2 | 17.5 | CCUG | (12) | 1.0 | 4.0 | 0 | 0 | [42] |
| Bovine PPPH | 342 | 32.7 | 16.4 | 26.0 | 24.9 | AAAA, AAAG, AGCU | (7) | 5.6 | 1.2 | 0 | 1.2 | [43] |
| Total 29 mRNA | 20157 | 27.8 | 20.6 | 22.4 | 29.2 | AAAA | (325) | 4.6 | 0.7 | 0.6 | 2.8 | |
| 119 sequences | 68571 | 25.8 | 23.5 | 24.0 | 26.6 | AAAA | (646) | 2.8 | 0.8 | 0.8 | 2.0 | [9] |

All published mRNA sequences for complete genes have been analyzed, as have all partial sequences of 50 or more codons. See text and legend of table 1 for description of 637ov3′ and 636vp1. The chicken ovalbumin coding sequence [3] appears here merely to show differences with the untranslated 3′ region (637ov3′), and is not included in the tally at the bottom of the table. Every other sequence given contains at least 4.0% of its total bases in runs of one base ($n > 3$). The expected amount of any base in runs of 4 or more is 1.28% of total bases in long sequences (over 300 bases) containing 25% of each base [8]. For most frequent quadruplet, each homogeneous quadruplet is calculated in all possible reading frames; thus a run of 5 bases includes 2 quadruplets, a run of 6 includes 3, etc. In all, 119 mRNA sequences were screened (the same sequences as in [9]). In some cases the coding region has not been perfectly delimited and the corresponding protein has not been sequenced sequences in which the amount of any base in runs of length greater than 3 nucleotides ($n > 3$) is at least 4.0% of total bases. The column totals demonstrate that in mRNA, A tends more than any other base to occur in runs. In 19 of the 29 mRNA in the table the most aggregated base is A. Note that although the overall base composition reveals more U than A in these messengers, $\%A_n$ (% of the total bases occurring as runs of adenine) is much higher than $\%U_n$. Also, in 12 of the 29 mRNA, AAAA appears as most frequent quadruplet while in none is UUUU the most frequent (UUUU is involved in two ties). Finally, 10 mRNA have greater overall frequencies for C, G or U than 636vp1 or 637ov3′ has for A, yet no one of the 10 shows $C_n$, $G_n$ or $U_n$ as high as $A_n$ in these two samples. The AAAA and $A_n$ contents of 636vp1 and 637ov3′ are very similar and exceed those of all other sequences in table 2.

It might be imagined that the high $A_n$ content of the mRNA in table 2 is simply a consequence of elevated lysine coding. This is not so. Of course the two phenomena of high $A_n$ in mRNA and frequent lysine coding are necessarily linked since AAA is a codon for this amino acid. However AAG, the other lysine codon, should occur as often as AAA if protein selection for lysine coding is the reason for the high $A_n$. Instead, AAA is found nearly twice as often (297-times) among the 6719 total codons in the 29 mRNA as AAG (157-times). In fact the proteins coded by these 29 mRNA sequences contain less lysine (6.8%) than does the average protein of Dayhoff (7.0%) [10]. Finally, an example where lysine coding *cannot* be related to high $A_n$ is 637ov3′ which has over 5-times as much $A_n$ as do the adjacent ovalbumin codons. Hence selection for lysine coding cannot generally explain poly(A) tendency.

In non-coding regions, as in codons, $\%A_n$ and $\%U_n$ are generally higher than $\%C_n$ or $\%G_n$ (table 3). Most samples in table 3 are associated with mRNA, in each such case the strand studied corresponds to an extension of the mRNA. Several untranslated sequences surpass the poly(A) tendency of 636vp1 and 637ov3′. As seen in table 3, however, in none of these cases is the CG/GC ratio extremely low as it is in 636vp1 or 637ov3′. The same is incidentally true of the yeast cytochrome *c* messenger of table 2 [11], which, although having AAAA as most frequent tetranucleotide and showing 7.4% $A_n$, also possesses a higher CG/GC ratio (6/11) than the two reference samples. Therefore, 636vp1 and 637ov3′ resemble each other much more than they do any of the table 2 or 3 sequences.

The *Xenopus laevis* ribosomal spacer DNA of table 3 is unusual. These sequences are different from any translated or untranslated sequence published by having a great quantity of C and G in runs. The spacers occur between 28 S and 18 S rRNA genes and are believed to have had a unique, rapid evolution [12].

This study has uncovered other interesting comparisons for the evolution of viruses. A surprising difference is found between the two sequences determined for the hepatitis B virus (HBV) surface antigen messenger [13,14]. The sequence (serotype ayw) of [14] contains 1.2% $A_n$, 1.9% $C_n$, 2.1% $G_n$ and 3.7% $U_n$ ($n > 3$). The same 675 nucleotides (initiator and terminator codons are excluded) in the sequence of [13] (serotype not given) show 2.4% $A_n$, 1.9% $C_n$, 2.1% $G_n$ and 3.1% $U_n$. Thirty-two, nearly 5%, of the bases vary between the two sequences and in 17 cases the change is in codon position I or II (16 amino acids differ). This contrasts with findings of a strong evolutionary pressure against mutation in coding regions between BKV wild-type and its variant (MM) [15]. Curiously, G varies <1/2 as often as any of the other bases. The number of appearances among the 64 different bases involved is A 19, C 20, G 7 and U 18. This suggests G is less mutable than the other bases (there is more G than A in these messengers [13,14]). Incidentally, in spite of the 32 base changes between the two sequences, a second reading frame remains entirely free of terminator codons throughout both sequences. This reinforces the suggestion [14,16] that the second frame is translated, perhaps to produce the DNA polymerase.

We can also compare complete genome sequences. A comparison of SV40 (monkey) and BKV (human) genome sequences [4,6] gives similar amounts of runs with each base between the two species (table 4). This is not true, however, of polyoma virus (mouse). It has been suggested (see, for example [7,17–22]) that polyoma and papova viruses have had a common ancestor, but the early polyoma region of 3013 nucleotides [18] contains only 1/3rd as much $A_n$ as the early part of SV40, as seen in table 4. Therefore, if polyoma and papova have derived from a common ancestor, which the organization of the genes in their genomes indeed favors, they have diverged radically in poly(A) tendency. Why would SV40 and BKV have acquired (or polyoma have lost) a high concen-

Table 3
Runs in untranslated sequences

| Untranslated sequence | No. bases | % | | | | CG/GC | [Ref.] |
|---|---|---|---|---|---|---|---|
| | | $A_n$ | $C_n$ | $G_n$ | $U_n$(T) | | |
| 636vp1 | 636 | 9.6 | 1.3 | 1.9 | 3.8 | 1/31 | [4] |
| 637ov3' | 637 | 8.5 | 0 | 1.4 | 1.9 | 0/23 | [3] |
| Ad 2 three 5'-leaders | 845 | 2.0 | 0.5 | 4.1 | 1.0 | | [44] |
| *E. coli* rL11, L1, L10 and L7 5' + 3' ends | 985 | 2.3 | 0.9 | 0.4 | 4.6 | | [39] |
| lipoprotein 5'-leader | 388 | 7.2 | 0 | 0 | 5.7 | 14/15 | [45] |
| rRNA promoters rrn E | 409 | 1.2 | 0 | 1.2 | 4.4 | | [46] |
| rrn A | 414 | 4.6 | 0 | 0 | 1.9 | | [46] |
| rrn D | 470 | 10.2 | 0.9 | 0 | 2.6 | 25/25 | [47] |
| rrn X | 476 | 6.7 | 0 | 0 | 1.9 | 27/31 | [47] |
| rrn B | 703 | 4.4 | 1.1 | 0.6 | 2.3 | | [48] |
| plasmid RK 526 | 520 | 3.3 | 0 | 0 | 6.9 | | [49] |
| Yeast 2 μ plasmid $H_2H_1R_2'$ zone | 1019 | 1.6 | 0 | 0.4 | 4.5 | | [50] |
| cytochrome *c* 5' + 3' ends | 522 | 2.3 | 1.9 | 0 | 6.9 | | [11] |
| mt petite mutant a1/1R/1 | 884 | 2.9 | 1.1 | 1.9 | 8.0 | | [51] |
| mt tRNA gene Ser | 320 | 5.3 | 1.2 | 0 | 5.0 | | [52] |
| gene Phe | 360 | 4.4 | 0 | 0 | 3.3 | | [52] |
| *D. discoideum* 5'-leader actin 5 | 502 | 20.9 | 0 | 0 | 23.9 | 1/4 | [53] |
| *Dros. melanogaster* satellite aDm 23-24 | 359 | 12.0 | 1.7 | 0 | 9.5 | 7/9 | [54] |
| *Bombyx mori* fibroin 5'-leader | 1001 | 8.8 | 0 | 0 | 3.3 | 26/28 | [40] |
| intron | 970 | 4.1 | 0.4 | 0.7 | 2.8 | | [40] |
| *Xenopus laevis* ribosomal RNA spacers | 1854 | 2.0 | 12.2 | 9.1 | 1.2 | 202/211 | [12] |
| Chicken conalbumin 5'-leader | 342 | 3.8 | 3.8 | 5.3 | 2.3 | | [24] |
| Rabbit β-globin introns | 699 | 1.1 | 1.1 | 1.3 | 10.4 | | [55,56] |
| Mouse β-globin intron β major | 653 | 1.1 | 2.0 | 1.2 | 4.6 | | [57] |
| intron β minor | 628 | 1.4 | 0.6 | 0.6 | 10.0 | | [57] |
| β minor 3' end | 371 | 4.6 | 2.2 | 2.2 | 3.2 | | [57] |
| Mouse α globin 5'-end | 404 | 0 | 1.0 | 4.4 | 1.0 | | [58] |

Since the shortest mRNA with 4.0% or more of total bases in runs of one kind of base contains 315 bases (TMV PSV 30K in table 2), a minimum length of 300 bases is taken for inclusion in this table. Published untranslated sequences appear in which runs ($n > 3$) of one kind of base account for at least 4.0% of total bases. When the sequence shown is associated with mRNA it corresponds to a continuation of the same strand. Fragments from the same general region of a genome have been combined to form some samples. For example, the *Xenopus laevis* ribosomal spacer sample combines sequences in fig.4,5 and 7 of [12]. The slime mold *Dictyostelium discoideum* 5'-leader for pDd actin 5 mRNA contains >90% A + T, which occurs largely in runs, as the table values for $A_n$ and $T_n$ indicate [53]. The CG/GC doublet frequency ratio is given for cases discussed in the text

tration of adenine runs? Could this aid in binding to primate nucleotide sequences, or in facilitating recombination? Whatever the reason, a functional distinction is indicated that perhaps can be detected by appropriate experiments. The cauliflower mosaic virus (CaMV) genome is also high in poly(A) tendency, but it does not have a low CG/GC ratio. Thus, these two indexes are not necessarily linked, as already suggested by table 3. It is curious that high concentrations of adenine runs occur both in ssRNA and dsDNA plant viruses (see tables 2 and 4).

The two polyoma early sequences [18,19] are rather alike. Aligning the A2 large plaque sequence in fig.2 of [19] and that (strain not given) in fig.4 of [18] shows 6 base changes, 8 insertions and 1 deletion in the latter. The differences are at positions 1216, 1664, 1666, 2035, 2038, 2042, 2194, 2482, 2491, 2500, 2607, 2609 and 2889 of the sequence in [19], in which, incidentally, an error occurs at position 2777 since the codon CGT is shown for glycine. This is <1/8th the variation seen above with the HBV surface antigen gene. Table 4 compares the sequences

Table 4
Base runs and CG/GC doublet frequency ratios in complete genome sequences having at least 4.0% of total bases in runs of one kind of base

| Genome sequence | No. bases | % ($n > 3$) | | | | CG/GC | [Ref.] |
|---|---|---|---|---|---|---|---|
| | | $A_n$ | $C_n$ | $G_n$ | $T_n$ | | |
| SV40 (Reddy et al.) | 5226 | 5.2 | 1.6 | 0.7 | 4.3 | 30/302 | [4] |
| SV40 (Fiers et al.) | 5224 | 5.8 | 0.6 | 1.6 | 3.7 | 27/263 | [5] |
| BKV (MM) | 4963 | 4.8 | 1.1 | 1.4 | 4.5 | 13/225 | [6] |
| BKV (Dun) | 5153 | 5.2 | 1.2 | 1.1 | 4.4 | 12/229 | [7] |
| py str. 3 large plaque | 5379 | 2.5 | 1.7 | 1.2 | 1.5 | 92/287 | [18,22] |
| py A2 strain | 5292 | 2.4 | 1.2 | 1.7 | 1.8 | 93/282 | [20] |
| HBV (SA strand) | 3182 | 1.1 | 1.6 | 1.9 | 2.6 | 99/161 | [14,16] |
| SV40 early | 2575 | 6.7 | 0.5 | 0.9 | 3.7 | 5/114 | [5] |
| SV40 late | 2649 | 4.9 | 0.6 | 2.3 | 3.6 | 22/149 | [5] |
| Polyoma early<br>strain 3 | 3013 | 1.9 | 1.9 | 0.9 | 1.7 | 54/169 | [18] |
| large plaque late | 2366 | 3.3 | 1.4 | 1.5 | 1.2 | 38/118 | [22] |
| CaMV | 8024 | 5.7 | 0.5 | 0.3 | 0.6 | 187/274 | [59] |
| Yeast plasmid | 6318 | 3.5 | 0.5 | 0.1 | 4.0 | 212/304 | [60] |

Values refer to one strand. Except for HBV and CaMV the genomes are believed to be entirely double stranded with the two strands fully complementary. Early and late regions of SV40 and polyoma are compared. SA, surface antigen

for SV40, BKV and polyoma determined by different groups and generally indicates less variation than in the HBV surface antigen case above.

Analysis of base runs in the entire HBV genome in [16], taking the same strand that codes the surface antigen, suggests that HBV resembles the polyoma early region more than it does papova viruses. The CG/GC doublet ratio values also support this idea, although the value is even higher for HBV than for polyoma (see table 4). No homology has been proposed in the case of HBV [13,14,16].

## 2. Conclusion

If the close $A_n$, AAAA and CG/GC contents in 636vp1 and 637ov3′ are not an accident, do they seriously indicate that the SV40 sequence may have originated as a fragment of the avian genome? Duplications are believed to have occurred in evolution of the ovalbumin gene [23,24]. Apparently, however, this gene has not been conserved in any mammal. Poly(A) tendency is found to a varying degree in several species and in different parts of the genome, as is evident in tables 2 and 3. The present analysis, therefore, is not proof of homology. Nevertheless, it must be recognized that if we accept the idea that viruses have originated as fragments of animal nucleotide sequence [25–28], the best candidate for ancestor of papova viruses among published sequences is still this untranslated ovalbumin 3′-end. Perhaps papova-like viruses exist in fowl or other birds. I hope the suggestion of possible horizontal transfer between avia and primates will stimulate further investigations.

These results strengthen my criticism of some well entrenched notions. The rareness of the CG dinucleotide in animal DNA continues to be evoked, but this rareness is much more pronounced in certain viruses [1]. The related idea that CG doublet rarity is a consequence of avoiding codons for arginine is false as well [1]. The CG doublet is indeed rare in some genes and in some entire genomes, as seen above. The reason it is rare remains unknown; I believe this relates to a protein-independent nucleic acid selection [1,2,9]. The same must be true of poly(A) tendency since it is a general sequence characteristic and is not confined to codons. The explanation for CG doublet avoidance could have a biophysical basis since CpG disrupts the normal helical structure of B-DNA [29]. However, no such basis with regard to runs of adenine has been suggested.

By the above indexes ($A_n$, AAAA and CG/GC contents) none of the mammalian sequences published, either coding or noncoding, shows the type of nucleotide organization found in papova viruses. Consequently, the origin of these viruses as a fragment of the mammalian genome seems unlikely. Of course only a few ppm of the mammalian genome have been sequenced.

## Acknowledgement

I thank Ray Wu and Francis Galibert for commenting on an earlier version of the manuscript.

## Note added in proof

Data for the *B. mori* sequence in table 2 should be ignored since the starting point taken for translation was too early in the sequence. Just which AUG initiates protein synthesis is still uncertain. Values for the untranslated *B. mori* sequences in table 3 are not affected.

## References

[1] Grantham, R. (1978) FEBS Lett. 95, 1–11.

[2] Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pavé, A. (1980) Nucleic Acids Res. 8, r49–r62.

[3] 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.

[4] 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.

[5] 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.

[6] Yang, R. C. A. and Wu, R. (1979) Science 206, 456–462.

[7] Seif, I., Khoury, G. and Dhar, R. (1979) Cell 18, 963–977.

[8] David, F. N. and Barton, D. E. (1962) in: Combinatorial Chance, pp. 85–101, Griffin, London.

[9] Grantham, R., Gautier, C. and Gouy, M. (1980) Nucleic Acids Res. 8, 1893–1912.

[10] Dayhoff, M. O. (1972) Atlas of Protein Sequence and Structure, p. D-335, Georgetown University Medical Center, Natl. Biomed. Res. Found., Washington DC.

[11] Smith, M., Leung, D. W., Gillam, S. and Astell, C. R. (1979) Cell 16, 753–761.

[12] Boseley, P., Moss, T., Mächler, M., Portmann, R. and Birnstiel, M. (1979) Cell 17, 19–31.

[13] Valenzuela, P., Gray, P., Quiroga, M., Zaldivar, J., Goodman, H. M. and Rutter, W. J. (1979) Nature 280, 815–819.

[14] Charnay, P., Mandart, E., Hampe, A., Fitoussi, F., Tiollais, P. and Galibert, F. (1979) Nucleic Acids Res. 7, 335–346.

[15] Yang, R. C. A. and Wu, R. (1979) Nucleic Acids Res. 7, 651–668.

[16] Galibert, F., Mandart, E., Fitoussi, F., Tiollais, P. and Charnay, P. (1979) Nature 281, 646–650.

[17] Howley, P. M. (1980) Nature 284, 124–125.

[18] Friedman, T., Esty, A., LaPorte, P. and Deininger, P. (1979) Cell 17, 715–724.

[19] Soeda, E., Arrand, J. R. and Griffin, B. E. (1979) Nucleic Acids Res. 7, 839–857.

[20] Soeda, E., Arrand, J. R., Smolar, N., Walsh, J. E. and Griffin, B. E. (1980) Nature 283, 445–453.

[21] Soeda, E., Maruyama, T., Arrand, J. R. and Griffin, B. E. (1980) Nature 285, 165–167.

[22] Deininger, P., Esty, A., LaPorte, P. and Friedmann, T. (1979) Cell 18, 771–779.

[23] Royal, A., Garapin, A., Cami, B., Perrin, F., Mandel, J. L., LeMeur, M., Brégegègre, F., Gannon, F., LePennec, J. P., Chambon, P. and Kourilsky, P. (1979) Nature 279, 125–132.

[24] Cochet, M., Gannon, F., Hen, R., Maroteaux, L., Perrin, F. and Chambon, P. (1979) Nature 282, 567–573.

[25] Temin, H. M. (1974) Annu. Rev. Gen. 8, 155–178.

[26] Temin, H. M. (1976) Science 192, 1075–1080.

[27] Subak-Sharpe, J. H., Elton, R. A. and Russell, G. J. (1974) in: Evolution in the Microbial World, 24th Symposium General Microbiology, pp. 131–150, (Carlile, M. J. and Stehel, J. J. eds) Cambridge University Press, London.

[28] Joklik, W. K. (1974) in: Evolution in the Microbial World, 24th Symp. Gen. Microbiol. (Carlile, M. J. and Stehel, J. J. eds) pp. 293–320, Cambridge University Press, London.

[29] Wang, A. H. J., Quigley, G. J., Kolpak, F. J., Crawford, J. L., Van Boom, J. H., Van der Marel, G. and Rich, A. (1979) Nature 282, 680–686.

[30] Guilley, H., Jonard, J., Kukla, B. and Richards, K. E. (1979) Nucleic Acids Res. 6, 1287–1308.

[31] Guilley, H. and Briand, J. P. (1978) Cell 15, 113–122.

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

[33] Shaw, D. C., Walker, J. E., Northrop, F. D., Barrell, B. G., Godson, G. N. and Fiddes, J. C. (1978) Nature 272, 510–515.
[34] Sugimoto, K., Oka, A., Sugisaki, H., Takanami, M., Nishimura, A., Yasuda, Y. and Hirota, Y. (1979) Proc. Natl. Acad. Sci. USA 76, 575–579.
[35] Van Wezenbeek, P. and Schoenmakers, J. G. G. (1979) Nucleic Acids Res. 6, 2799–2818.
[36] Sauer, R. (1978) Nature 276, 301–302.
[37] Scherer, G. (1978) Nucleic Acids Res. 5, 3141–3156.
[38] Marcoli, R., Iida, S. and Bickle, T. A. (1980) FEBS Lett. 110, 11–14.
[39] Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H. and Dennis, P. P. (1979) Proc. Natl. Acad. Sci. USA 76, 1697–1701.
[40] Tsugimoto, Y. and Suzuki, Y. (1980) Cell 18, 591–600.
[41] Seidman, J. G., Leder, A., Nau, M., Norman, B. and Leder, P. (1978) Science 202, 11–17.
[42] Heindell, H. C., Liu, A., Paddock, G. V., Studnicka, G. M. and Salser, W. A. (1978) Cell 15, 43–54.
[43] Kronenberg, H. M., McDevitt, B. E., Majzoub, J. A., Nathans, J., Sharp, P. A., Potts, J. T. jr and Rich, A. (1979) Proc. Natl. Acad. Sci. USA 76, 4981–4985.
[44] Askusjärvi, G. and Pettersson, U. (1979) J. Mol. Biol. 134, 143–158.
[45] Nakamura, K. and Inouye, M. (1979) Cell 18, 1109–1117.
[46] De Boer, H. A., Gilbert, S. F. and Nomura, M. (1979) Cell 17, 201–209.
[47] Young, R. A. and Steitz, J. A. (1979) Cell 17, 225–234.
[48] Csordas-Toth, E., Boros, I. and Venetianer, P. (1979) Nucleic Acids Res. 7, 2189–2197.
[49] Stalker, D. M., Kolter, R. and Helsinki, D. R. (1979) Proc. Natl. Acad. Sci. USA 76, 1150–1154.
[50] Hindley, J. and Phear, G. A. (1979) Nucleic Acids Res. 7, 361–375.
[51] Gaillard, C., Strauss, F. and Bernardi, G. (1980) Nature 283, 218–220.
[52] Miller, D. L., Martin, N. C., Dinh Pham, H. and Donelson, J. E. (1979) J. Biol. Chem. 254, 11735–11740.
[53] Firtel, R. A., Timm, R., Kimmel, A. R. and McKeown, M. (1979) Proc. Natl. Acad. Sci. USA 76, 6206–6210.
[54] Hsieh, T.-S. and Brutlag, D. (1979) J. Mol. Biol. 135, 465–481.
[55] Van Ooyen, A., Van den Berg, J., Mantei, N. and Weissman, C. (1979) Science 206, 337–344.
[56] Hardison, R. C., Butler, E. T. III, Lacy, E., Maniatis, T., Rosenthal, N. and Efstratiadis, A. (1979) 18, 1285–1297.
[57] Konkel, D. A., Maizel, J. V. jr and Leder, P. (1979) Cell 18, 865–873.
[58] Nishioka, Y. and Leder, P. (1979) Cell 18, 875–882.
[59] Franck, A., Guilley, H., Jonard, G., Richards, K. and Hirth, L. (1980) Cell 21, 285–294.
[60] Hartley, J. L. and Donelson, D. E. (1980) Nature 286, 860–864.