An oligopurine sequence bias occurs in eukaryotic viruses

Anne M. Beasty and Michael J. Behe

Department of Chemistry, Lehigh University, Seeley Mudd Building no.6, Bethlehem, PA 18015, USA

Received January 7, 1988; Accepted January 21, 1988

ABSTRACT

Twenty four DNA and RNA viral nucleotide sequences, comprising over 346 kilobases, have been analyzed for the occurrence of strings of contiguous purine or pyrimidine residues. On average strings greater than or equal to 10 contiguous purines or pyrimidines are found three and a half times more frequently than would be expected for a random distribution of bases. Detailed analysis of the 172 kilobase Epstein-Barr viral sequence shows that the bias in favor of contiguous purine residues increases with the length of the purine string. These findings are similar to those seen for genomic DNA from higher eukaryotes (1). In contrast no overrepresentation of oligopurine or oligopyrimidine strings is observed in 52 kilobases from eight bacteriophage and *E. coli* DNA sequences.

INTRODUCTION

The genomic DNA of higher eukaryotes contains a much larger number of base pairs than is required to code for the protein products of the genome. It is estimated that only about 10% of the DNA is found in coding sequences (2). The functions of the remaining 90% of the genome are ill-defined, although they certainly include control sequences and structural components.

A large number of repetitive sequences are found in eukaryotes. Simple repeating sequences and short and long interspersed repetitive elements are relatively abundant (2), sometimes making up as much as 50% of the total DNA of an organism. This has led to speculation that some of the non-coding regions really have no specific function, and remain because there is not sufficiently strong evolutionary pressure on the organism to rid itself of extraneous DNA sequences (3,4).

In a recent paper (1) we analyzed the 67 kilobase DNA sequence of the human β -globin region and over 80 kilobases of other human DNA sequences for the occurrence of strings of

contiguous purine or pyrimidine residues. Tracts of ten or more contiguous purine or pyrimidine residues were found four times more frequently than would be expected for a random distribution of bases so that a long string occurs at an average of every 250 base pairs and accounts for a total of ~5% of the DNA. Since the frequency of occurrence of these strings corresponds closely to the nucleosomal repeat unit and since some synthetic polydeoxynucleotides in which one strand is composed completely of purine residues can not form normal nucleosomal structures (5-8), we speculated that these elements may play a role in influencing the structure of chromatin *in vivo*. However, since the genomes of higher eukaryotes tolerate DNA sequences with no apparent function it might be possible that the overabundant oligopurine strings accumulated serendipitously, to no real purpose.

In this paper we examine the sequences of some eukaryotic viruses for strings of contiguous purines and pyrimidines. The viral genomes we consider occur in the eukaryotic nucleus, either free or integrated into the host DNA as proviral sequences. There, at some point, they probably form nucleoprotein structures similar to the host DNA. However, unlike genomic eukaryotic DNA, viruses are under considerable selective pressure to minimize their genome size. Additionally, and again unlike genomic DNA, viral sequences consist predominantly of coding regions. Thus it might be expected that extraneous, non-functional elements would be purged from viruses.

The results presented here show that in viral nucleotide sequences strings of contiguous purine or pyrimidine residues equal to or exceeding ten bases in length are found, on average, at a frequency more than three and a half times greater than expected for random occurrence of bases, similar to their overabundance in genomic eukaryotic DNA. This observation supports our hypothesis (1) that oligopurine strings do have a positive role in the function of the eukaryotic genome.

METHODS

DNA sequences found in the Genbank genetic sequence data bank (ACM, ACY, and MLFRO are from Release 48.0, February 1987; all

other files are from Release 40.0, February 1986) were searched using a microcomputer with commercial sequence analysis programs (International Biotechnologies, Inc.) written by Pustell & Kafatos (9-11). A description of the sequences searched, the size of the sequence files, and the Genbank file designations are shown in Tables 1 & 2.

Since, of course, oligopyrimidine strings are the complement of oligopurine strings all sequence files were searched for both regions of contiguous purines and regions of contiguous pyrimidines. The Epstein-Barr virus sequence (12) file and the prokaryotic sequence files were searched for occurrences of the sequences $\text{-Py-(Pu)}_n\text{-Py-}$ and $\text{-Pu-(Py)}_n\text{-Pu-}$, with n ranging from 1 to 15. The other, smaller eukaryotic files were searched simply for strings of 10 to 15 contiguous purines or pyrimidines ($n=10$ to 15). All sequences were then searched for occurrences of $\text{-(Pu)}_{16}\text{-}$ and $\text{-(Py)}_{16}\text{-}$, which yields all strings of 16 or greater.

Expected frequencies of $\text{-Py-(Pu)}_n\text{-Py-}$ strings for a random distribution of bases can be calculated in several different ways. In this paper we take the simplest approach. Since all of the sequences we analyzed exist at some point as double stranded DNA (the single stranded RNA viruses are all reverse-transcribed and integrated as double stranded DNA into the host genome and the single stranded virus (AAV2) and phages replicate as double stranded DNA) then, in the double stranded form, exactly 50% of the residues of the DNA are purines and 50% are pyrimidines. Thus the expected frequencies of $\text{-Py-(Pu)}_n\text{-Py-}$ were calculated as follows: (expected number of occurrences) = (number of bases in data set) $\times 0.5^{(n+2)}$. Expected frequencies of $\text{-Pu-(Py)}_n\text{-Pu-}$ were calculated analogously. This method takes into account neither di- or tri- nucleotide frequencies nor strand asymmetry. However, taking such considerations into account introduces the assumption that they were not "intentionally" designed to affect the number of oligopurine strings or that they are simply reflections of the occurrence of such strings. Since these may not be valid assumptions we confine ourselves here to the simple method of calculation described above. Fortunately the results are changed very little in any event and some sample calculations are included in the Results section to demonstrate this.

Table 1. Genbank files of eukaryotic sequences searched.

Genbank file designation	size of file (bases)	Description of file
<u>DNA viruses</u>		
LPV	5,270	Lymphotropic papovavirus
BKVM	4,963	Human papovavirus BK, (MM strain)
JCV	5,130	JC polyomavirus
AAV2	4,675	Adeno associated virus 2
BPV1	7,945	Bovine papillomavirus type 1
SV40	5,243	Simian virus 40
EBV	172,282	Epstein-Barr virus, B95-8 strain
PYA2	5,297	Polyoma virus strain a2
DHBVG	3,021	Duck hepatitis B virus
CRPCG	7,868	Cottontail rabbit papillomavirus
AD2	35,937	Adenovirus, type 2
HPV16	7,904	Human papillomavirus, type 16
<u>RNA viruses</u>		
HTLV1PROP	9,067	Human T-cell leukemia virus
ACM	3,779	Avian myelocytomatosis virus MC29
ACY	3,718	Avian sarcoma virus Y73
EIAGP	4,799	Equine infectious anemia virus
SSV	5,779	Simian sarcoma virus
MMSV124	5,833	Moloney murine sarcoma virus
AIDLAVA	9,769	AIDS/lymphadenopathy retrovirus
AKV	8,374	AKV murine leukemia virus
MLFRO	6,296	Friend spleen focus-forming virus
BLV	8,714	Bovine leukemia virus
RSCCG	9,625	Rous sarcoma virus
FSV	4,788	Fujinami sarcoma virus

RESULTS

The number of occurrences of the sequences $-\text{Py}-(\text{Pu})_n-\text{Py}-$ and $-\text{Pu}-(\text{Py})_n-\text{Pu}-$ is plotted in Fig. 1 versus n , the number of contiguous purine or pyrimidine residues, for the sum of the prokaryotic DNA sequences listed in Table 2. The solid line in the figure is not drawn to fit the data but passes through the predicted frequencies for the random occurrence of sequences of

Table 2. Genbank files of prokaryotic sequences searched.

Genbank file designation	size of file (bases)	Description of file
<u>Phage sequences</u>		
IKECG	6,883	Bacteriophage IKE
PHIX174	5,386	Bacteriophage ϕ X174
FD	6,408	Bacteriophage fd, strain 478
G4	5,577	Bacteriophage G4
<u><i>E. coli</i> sequences</u>		
ECOTRP	7,335	<i>E. coli</i> tryptophan operon
ECOACE	7,740	<i>E. coli</i> ace operon
ECOAMPCFER	5,482	<i>E. coli</i> frd operon
ECOLAC	7,477	<i>E. coli</i> lactose operon

the types that were searched for. As seen the data fit the predicted line very closely. The standard deviation of a random sequence from the calculated value is shown by the dashed curves. The sum of all prokaryotic sequences was used in order to provide a large data base. However, each individual gene shows the same

Figure 1. Number of $-Py-(Pu)_n-Py-$ plus $-Pu-(Py)_n-Pu-$ sequences versus n , the number of contiguous purine or pyrimidine residues, for the sum of the prokaryotic DNA sequences listed in Table 2.

Table 3. Ratios of observed/expected number of strings consisting of ten or more contiguous purines or pyrimidines for prokaryotic sequences.

Phage sequences				<i>E. coli</i> sequences			
File	Exp'd	Obs'd	Ratio	File	Exp'd	Obs'd	Ratio
IKECG	7	15	2.2	ECOTRP	7	10	1.4
PHIX174	5	5	1.0	ECOACE	7	8	1.1
FD	6	5	0.8	ECOAMPCFER	5	4	0.8
G4	5	3	0.6	ECOLAC	7	5	0.7
Average ratio = 1.1				Average ratio = 1.0			

behavior, albeit with a bit more scatter. Table 3 shows the expected and observed number of occurrences of strings of 10 or more contiguous purines and pyrimidines broken down for the individual prokaryotic sequence files. All of the *E. coli* and bacteriophage sequences, with the exception of IKECG, have ratios very close to one, and the average ratio for the bacteriophages and *E. coli* sequences as a group is 1.0.

A similar plot is shown in Fig. 2 for the sequence of the Epstein-Barr virus. For this eukaryotic virus, however, the data do not fall along the predicted line, but deviate in a positive

Figure 2. Number of $-\text{Py}-(\text{Pu})_n-\text{Py}-$ plus $-\text{Pu}-(\text{Py})_n-\text{Pu}-$ sequences versus n , the number of contiguous purine or pyrimidine residues, for the DNA sequence of the Epstein-Barr virus.

Table 4. Ratios of observed/expected number of strings consisting of ten or more contiguous purines or pyrimidines for eukaryotic sequences.

DNA viruses				RNA viruses			
<u>File</u>	<u>Exp'd</u>	<u>Obs'd</u>	<u>Ratio</u>	<u>File</u>	<u>Exp'd</u>	<u>Obs'd</u>	<u>Ratio</u>
LPV	5	32	6.4	HTLV1PROP	9	46	5.1
BKVM	5	27	5.4	ACM	4	20	5.0
DHBVG	3	16	5.3	ACY	4	17	4.3
JCV	5	25	5.0	EIAGP	5	21	4.2
AAV2	5	17	3.4	SSV	6	25	4.2
BPV1	8	27	3.4	MMSV124	6	24	4.0
PYA2	5	17	3.4	AIDLAVA	10	37	3.7
SV40	5	16	3.2	AKV	8	29	3.6
EBV	168	508	3.0	MLFRO	6	19	3.2
CRPCG	8	16	2.0	BLV	9	28	3.1
AD2	35	66	1.9	RSCCG	9	26	2.9
HPV16	8	10	1.3	FSV	5	13	2.6
Average ratio = 3.6				Average ratio = 3.8			

direction as the length of contiguous purines and pyrimidines becomes greater. Counted occurrences of purine strings first exceed expected occurrences for $n=5$. The deviation increases as the length of the purine string increases. At $n=10$ the ratio of expected/observed is 2.1 and increases to 3.9 for $n=14$, with an overall ratio of 3.0 for all strings of 10 or greater.

Table 4 shows the expected and observed number of strings of ten or more contiguous purines or pyrimidines for the Epstein-Barr virus, eleven other eukaryotic DNA viruses (ten are double stranded and one, AAV2, is single stranded), and twelve single stranded RNA eukaryotic viruses. The ratio of observed to expected number of strings varies for both RNA and DNA viruses, from a low of 1.3 for HPV16 to a high of 6.4 for LPV, with an average of 3.7 for the eukaryotic viruses as a group. (The averages shown in Table 3 are simple number averages of the

unweighted values for each member of a given category.)

On the mononucleotide level the expected number of strings of purines of a particular length can be calculated as: expected number of strings = (number of base pairs in data set) x (frequency of pyrimidines)² x (frequency of purines)ⁿ (1). If purines and pyrimidines are equally distributed between the two strands of double stranded DNA then this reduces to the formula presented in the Methods section. However, if there is an asymmetric distribution of purines and pyrimidines between the two strands then there is an increased expected number of strings. If one strand is 53% purines then 1.1 times the number of strings with ten or more contiguous purines or pyrimidines are expected than for the case of equal distribution of purine on both strands. With strands containing 56% or 59% purines the expected number of long strings goes to 1.5 and 2.2 times the number expected for an equal distribution on both strands. Fortunately, only 9 out of the 32 sequences listed in Tables 3 & 4 have strand asymmetries greater than 53%. When strand asymmetry is taken into account the average ratio for observed to expected number of strings for DNA viruses goes from 3.6 to 3.4, for RNA viruses it changes from 3.8 to 3.2, and the ratio remains the same at 1.0 for the prokaryotic sequences. Thus the results are not greatly affected by this consideration.

Table 5 shows the ratio of observed to expected number of strings broken down into coding and non-coding regions for the eukaryotic Epstein-Barr, adeno-, and SV40 viruses. EBV contains ~56,000 base pairs of non-coding sequences. There are 4.6 times the number of expected purine or pyrimidine strings of ten or greater in length in these non-coding regions. For the coding regions of the Epstein-Barr virus (listed as open reading frames in the Genbank annotation) the ratio is 2.4. The SV40 DNA sequence has an overall ratio similar to that of the Epstein-Barr virus, but the break-down into coding and non-coding regions is reversed, -for SV40 the non-coding region has a smaller ratio than the coding region. However, since SV40 is 88% coding there are only about 600 base pairs of non-coding DNA which includes only one string of greater than 10 base pairs, so practically all long strings are in coding regions in SV40. For adenovirus, which

Table 5. Ratios of observed/expected number of strings consisting of ten or more contiguous purines or pyrimidines for coding versus non-coding regions of several eukaryotic viruses.

virus	% coding sequence	Ratio		
		coding	non-coding	total
EBV	69	2.4	4.6	3.0
SV40	88	3.3	1.7	3.2
AD2	74	1.8	2.0	1.9

contains about 9000 base pairs of non-coding sequences and 27000 base pairs of coding sequences, the ratios are very similar for the two regions.

DISCUSSION

Oligopurine strings in viral DNA

Eukaryotic cells contain much more DNA than is required to code for the protein products of their genomes (2). Thus it has been questioned whether sequences such as short and long interspersed repetitive elements have any specific function in the cell, or if they simply accumulate over time by various mechanisms and there is not sufficient evolutionary pressure on the organism to rid itself of extraneous DNA sequences (3,4).

Recently we showed (1) that the DNA sequence of the human β -globin gene region, and probably the DNA of most higher eukaryotes, is strongly biased in favor of long strings of contiguous purine or pyrimidine residues. Since strings of 10 or more contiguous residues were seen to compose ~5% of the total DNA and to occur on average once every 250 base pairs, close to the nucleosomal repeat distance, and since some synthetic polydeoxynucleotides with one strand composed entirely of purines fail to form nucleosomes *in vitro* (5-8), we speculated that these elements may play a basic role in influencing the structure of chromatin *in vivo*. We showed that the oligopurine strings were not repetitive sequences and thus probably did not arise from any simple mechanism for duplicating DNA sequences. Nonetheless, the question remained whether these sequences are constructive

elements, performing some function or functions, or if they arose serendipitously and are essentially non-functional.

To attempt an answer to that question we decided to examine eukaryotic DNA sequences that are under selective pressure to keep their genomes small and functional, -the viruses. Our reasoning was that if overabundant oligopurine and oligopyrimidine strings did not serve a positive role in eukaryotic DNA then they would have been eliminated in the viruses to make room for functional sequences. However, if such sequences are necessary for the functioning of double stranded DNA in a eukaryotic nucleus, then they might be found in DNA viruses and might also be reflected in the sequences of RNA viruses that integrate into the host genome and exist there as double stranded DNA.

As seen in Figure 2 and Table 4 long strings (≥ 10 base pairs) of oligopurine and oligopyrimidine residues are overrepresented in almost all of the eukaryotic DNA and RNA viruses examined. On average there are over three and a half times as many such strings than expected. This is comparable to, although slightly less than, eukaryotic genomic DNAs (1). There is more variability in the viral sequences than in the DNA of higher eukaryotes, but only 2 out of 24 viruses have ratios of observed to expected number of strings less than 2.0. The ratio does not correlate with the size of the viral DNA. A comparatively large virus, adenovirus, has a lower ratio than almost all smaller viruses.

In contrast, with one exception, the prokaryotic bacteriophages and *E. coli* DNA sequences all have ratios very close to 1.0 (Table 3). Thus we have shown that the existence of a sequence bias in favor of long strings of purine or pyrimidine residues does not correlate with percentage of coding regions, amount of selective pressure, or whether information is carried as DNA or RNA. The only correlation we have been able to see is with eukaryotic vs. prokaryotic sequences. We believe that these results greatly strengthen our thesis that such elements have a positive role in the expression of eukaryotic DNA function.

Implications for the genetic code

The question arises, since the overproduction of oligopurine and oligopyrimidine strings occurs mostly in coding sequences in

the viral genomes discussed above, whether an undue burden is placed on the genetic code to produce these strings while simultaneously coding for functional proteins. It seems unlikely that this is the case. Of the 20 amino acids 3 *must* be designated by contiguous purines or pyrimidines (phe, lys, glu), and another five codons (leu, ser, pro, arg, gly) *can* be coded for by contiguous purines or pyrimidines. Therefore the interiors of long strings in coding regions can consist of either phe, leu, ser, and pro codons or, depending on which strand is the sense strand, lys, glu, arg, and gly codons. Furthermore, the final two bases at both ends of oligopurine or oligopyrimidine strings can be contributed by codons for other amino acids. Four other amino acid codons (ile, val, thr, ala), although unable to consist of contiguous purine or pyrimidine residues, can have either a purine or pyrimidine in the wobble position and can therefore adjust their codon to lengthen an existing string by one or two nucleotides. Met, tyr, trp, his, cys, gln, asn, and asp codons can not be adjusted to change the occurrence of purines or pyrimidines. It is interesting to note that the first five of these are the least abundant amino acid residues in proteins (13). It seems, therefore, that there is sufficient flexibility in the redundancy of the genetic code to simultaneously accomodate an overabundance of long purine or pyrimidine strings while coding for functional proteins.

ACKNOWLEDGEMENTS

This work was supported by grant GM36343 from the National Institutes of Health and grant DMB-8510719 from the National Science Foundation. M.J.B. is a recipient of Research Career Development Award CA01159 from the National Institutes of Health.

REFERENCES

1. Behe, M. J. (1987) *Biochemistry* 26, 7870-7875.
2. Lewin, B. (1985) in *Genes* 2 pp 293-295, John Wiley & Sons, New York.
3. Doolittle, W. F. & Sapienza, C. (1980) *Nature* 284, 601-603.
4. Orgel, L. E. & Crick, F. H. C. (1980) *Nature* 284, 604-607.
5. Bryan, P.N., Wright, E.B., Hsie, M.H., Olins, A.L. & Olins, D.E. (1978) *Nucleic Acids Res.* 5, 3603-3617.
6. Rhodes, D. (1979) *Nucleic Acids Res.* 6, 1805-1816.
7. Simpson, R.T. & Kunzler, P. (1979) *Nucleic Acids Res.* 6, 1387-1415.

8. Kunkel, G.R. & Martinson, H.G. (1981) *Nucleic Acids Res.* 9, 6869- 6888.
9. Pustell, J. & Kafatos, F. C. (1982a) *Nucleic Acids Res.* 10, 51-59.
10. Pustell, J. & Kafatos, F. C. (1982b) *Nucleic Acids Res.* 10, 4765-4782.
11. Pustell, J. & Kafatos, F. C. (1984) *Nucleic Acids Res.* 12, 643-655.
12. Baer, R., Bankier, A. T., Biggin, M. D., Deininger, P. L., Farrell, P. J., Gibson, T. J., Hatfull, G., Hudson, G.S., Satchwell, S. C., Seguin, C., Tuffnell, P. S., & Barrell, B. G. (1984) *Nature* 310, 207-211.
13. Schulz, G. E. & Schirmer, R. H. (1985) in *Principles of Protein Structure* p 2, Springer-Verlag, New York.