Statistical significance of symmetrical and repetitive segments in DNA

Gary R. Day* and R. D. Blake

Departments of Mathematics and Biochemistry, University of Maine, Orono, ME 04473, USA

Received 27 July 1982; Revised and Accepted 18 October 1982

Abstract

Methods of computer analysis for the recurrence of symmetrical and repetitive elements in large numbers of DNA sequences are described, together with derivations of appropriate quantitative criteria for the evaluation of the statistical significance of these elements in DNAs of different base composition. Examples of some extraordinary variations in the occurrence of symmetrical and repetitive elements are provided, many of which are new.

Special consideration is devoted to a determination of the statistical significance of a two-fold palindrome at the origin of replication. A computer search of 14 independently determined DNA sequences containing an origin of replication locus indicates each contains a large two-fold palindrome. The average length of this palindrome is 28 ± 6 base pairs, of which 22 contribute to the palindromic symmetry. The probability of occurrence of such a palindrome is only $1/26000$, while the probability of occurrence in all 14 different species is $(1/26000)^{14}$.

INTRODUCTION

Interest in the occurrence of various symmetrical and repetitive elements in DNA sequences has led to numerous schemes for their function and origin. The occurrence of large palindromic sequences with two-fold symmetry at the origin of replication of many genomes (eg., 1-6) is an especially intriguing example, since it would appear to represent a clue to the molecular mechanism of initiation of replication. Also, the recurrence of short sequences represents a possible clue to the mechanism of shifts in gene organization and duplication (7). It is, therefore, of interest to determine whether the occurrence or lack of occurrence of such elements are the consequence of statistical accident and, therefore, whether further investigation of their biological or evolutionary significance might be warranted.

In this report we describe efficient methods of computer analysis for symmetrical and repetitive elements in DNA, together with a description of appropriate quantitative (statistical) criteria to determine whether such

elements differ significantly in frequency of occurrence from that expected from chance occurrence; eventually to relate such differences to particular thermodynamic properties of the base pairs, or to the properties of enzymes catalyzing translation and the replication, recombination and repair of DNA.

METHODS OF ANALYSIS

Oligonucleotide frequencies: A) Within a sequence. Recurring oligonucleotide sequences (tuples) are determined with the routine NTUPLE, which analyzes any sequence for frequencies of tuples of any length n , resulting in tables of 4^n lines in length. The observed tuplet frequency is determined from a simple count, while the relative frequency is obtained from the ratio of observed frequency to sequence length. The expected relative frequency is calculated as the product of the observed relative frequencies of the bases in the primary strand (or complementary strand if that is selected), while the expected frequency is the product of expected relative frequency and length.

B) Among different sequences. The observed frequencies of matching oligonucleotide sequences in two different DNAs is determined with routine COMPARE. Let the fractional concentration of A be $\underline{a}$; of T, $\underline{t}$; etc., and let the base contents in the strand of interest in the first sequence be subscripted by $\underline{1}$ and of the second sequence by $\underline{2}$. Then the probability of finding a single residue match at any given point in segments of both sequences is given by

$$P_{1,2} = a_1 a_2 + c_1 c_2 + g_1 g_2 + t_1 t_2 \quad (1)$$

while the probability for an exception is $(1-p_{1,2})$. The a priori probability of finding two identical segments of length $\underline{n}$ with $\underline{s}$ matching pairs is then given by

$$p_s = n! p_{1,2}^s (1-p_{1,2})^{n-2} / s! (n-s)! \quad (2)$$

and for finding at least $\underline{m}$ out of $\underline{n}$ matching segments

$$p_m = \sum_{s=m}^n n! p_{1,2}^s (1-p_{1,2})^{n-s} / s! (n-s)! \quad (3)$$

For a DNA of length L the mean is $L_1 L_2 p_s$. Let $J = L_1 L_2$, then the probability of finding N identical segments is

$$p_N = J! p_s^N (1-p_s)^{J-N} / N! (J-N)! \quad (4)$$

while the probability of finding at least N identical segments is

$$P_N + 1 - \sum_{i=0}^{N-1} J! p_s^i (1-p_s)^{J-i} / i! (J-i)! \quad (5)$$

Two-fold palindromes. We define a palindrome as any segment of length $2n$ or $2n+1$ bp exhibiting from zero to n symmetries. To obtain two-fold symmetry at any point a residue in the primary strand of the symmetrical left-half palindrome must be matched by the complement residue in the primary strand of the right-half palindrome. Analysis for the observed frequency of both $2n$ and $2n+1$ -type two-fold palindromes is performed with the routine TWOFLD. Attempts to make a priori calculation of the expected frequency of two-fold palindromes generally recognize the number of both symmetrical and nonsymmetrical or hyphenated residues in the palindrome (8). Base composition is equally important in determining expected frequencies of two-fold palindromes.

If the fractional concentration of A in the primary strand is a ; of T, t ; etc., then the probability of achieving symmetry by matching an A in the left-half palindrome with a T in the right-half palindrome is $(a \cdot t)$, etc. For all possible symmetries in a two-fold palindrome to occur simultaneously we have $(a \cdot t) + (t \cdot a) + (g \cdot c) + (c \cdot g)$, so that the probability of finding two-fold symmetry at a given point in the palindrome is given by

$$P_{tf} = 2(a \cdot t + g \cdot c) \quad (6)$$

The probability for the occurrence of a palindrome of half-length n with axis at a particular position is then given by

$$p_s + (n! p_{tf}^s (1-p_{tf})^{n-s}) / (n-s)! s! \quad (7)$$

where s is in the number of base pairs in the palindrome exhibiting perfect symmetry. The expected frequency, λ , of finding a palindrome in a DNA of length L is just $2Lp_s$; where the factor of two is in recognition of palindromes with an axis of symmetry between base pairs, that is, palindromes of length $2n$, plus those with axis over a base pair of length $2n+1$. Further distinctions of symmetries as purine-pyrimidine in the two-fold case or purine-purine in the mirror case are not considered explicitly, the appropriate equations may be derived from reference to Brezinski's paper (8) and equations 6 and 7.

The probability that at least m out of the n positions in a palindrome will be symmetrical is given by

$$P^m = \sum_{s=m}^n (n! p_{tf}^s (1-p_{tf})^{n-s}) / s! (n-s)! \quad (8)$$

while the probability of finding N palindromes in a DNA of length L is given by the binomial distribution function

$$P_N = L! p_s^N (1-p_s)^{L-N} / N! (L-N)! \quad (9)$$

Usually L is large and computation of factorials larger than, say, 50 generally will not be possible; in which case the Poisson distribution is more appropriate

$$P_N = \lambda^N \cdot e^{-\lambda} / N! \quad (10)$$

The probability of obtaining at least N palindromes is represented by a cumulative binomial distribution, and since L is large and p_s is small, will be closely approximated by a cumulative Poisson

$$P_N = 1 - \sum_{i=0}^{N-1} \lambda^i \cdot e^{-\lambda} / i! \quad (11)$$

Palindromes with mirror symmetry. Analysis for the observed frequency of both $2n$ and $2n+1$ -type palindromes exhibiting mirror symmetry is performed with routine MIRROR. The probability of finding symmetry at a given point in a palindrome with mirror symmetry is given by

$$p_{\text{mir}} = a^2 + c^2 + g^2 + t^2 \quad (12)$$

where a , c , g , and t are the fractional content of A, C, G, T in a DNA of length L . The formulas above, eqns. 7-11 are then used with substitution of p_{mir} for p_{tf} .

Repeated sequences. Repeated sequences exhibit translational symmetry and, therefore, their occurrence in DNA represent a logical extension of analyses for two-fold and mirror palindromes. Since the length of the translational axis cannot be a priori specified in this case, our analysis with routine MATCH searches for all segments of matching sequence in the same strand. The probability of finding repetition at a given point in either contiguous or noncontiguous repeating segments is given by

$$p_r = a^2 + c^2 + g^2 + t^2 \quad (13)$$

The probability of finding a repeated sequence of length n with s symmetries is given by

$$p_s = n! p_r^s (1-p_r)^{n-s} / s! (n-s)! \quad (14)$$

while the probability of a repeated segment with at least m symmetries is

$$p_m = \sum_{s=m}^n n! p_r^s (1-p_r)^{n-s} / s! (n-s)! \quad (15)$$

The expected frequency of pairs of identical segments is then $L(L-1)p_s/2$.

Let

$$J = L(L-1)/2 \quad (16)$$

then the probability of N pairs is given by

$$p_N = J! p_s^N (1-p_s)^{J-N} / N! (J-N)! \quad (17)$$

The probability of finding at least N pairs is then

$$P_N = 1 - \sum_{i=0}^{N-1} J! p_s^i (1-p_s)^{J-i} / i! (J-i)! \quad (18)$$

Inverted Sequences. Inverted repeating sequences differ from two-fold palindromes in not having a readily recognizable axis of symmetry, since the repeat is not required to be contiguous but may exist anywhere in the complementary strand. Analysis for their occurrence in DNA is performed with the routine INVERT.

The probability of finding an inverted repeat at a given point in two segments is the same as for two-fold palindromes (eqn. 6) except that the segments are not contiguous. The distributions derived above for repeating sequences then hold for inverted repeats, except that

$$J = L(L+1)/2 \quad (19)$$

Since L will generally be large and p_{tf} will be small, the cumulative binomial distribution can be substituted by a Poisson distribution, where the mean is given by $L(L+1)p_s/2$.

RESULTS AND DISCUSSION

Oligonucleotide frequencies. The evolution of sequences includes both single base changes as well as the recombination of fragments, consequently we have begun a study of oligonucleotide frequencies in sequences from *E. coli* and the yeast *S. cerevisiae*. Our analysis presently involves more than 100kbp from just these two species. Not surprisingly we find extraordinary deviations in the observed frequencies of certain lengthy tuples from that expected by chance occurrence, however, only some of these deviations can be explained by the products of the observed frequencies of smaller, component tuples.

In our analysis of duplet frequencies we find significant deviations or "asymmetries" (9) from frequencies expected from random occurrence.

Table 1. Duplet Statistics for the Primary Strand of 20 DNA Sequences* of 29,655 bp from E. Coli.

Duplet (Rank Order)	Frequency	Relative Frequency	Expected Frequency	Expected Relative Frequency	Percent Difference
AA	2294	0.07736	1863.9	0.06285	+23.08
GC	2268	0.07648	1887.6	0.06365	+20.15
TG	2216	0.07473	1893.6	0.06385	+17.03
CG	2126	0.07169	1887.6	0.06365	+12.63
TT	2037	0.06869	1893.6	0.06385	+ 9.81
AT	1954	0.06589	1841.5	0.06210	+ 6.11
CA	1821	0.06141	1769.3	0.05966	+ 2.92
GA	1897	0.06397	1908.2	0.06435	- 0.59
TC	1728	0.05827	1762.2	0.05942	- 1.94
GT	1743	0.05878	1893.6	0.06385	- 7.95
CC	1639	0.05527	1785.2	0.06020	- 8.19
CT	1617	0.05453	1762.2	0.05942	- 8.24
GG	1803	0.06080	2021.2	0.06816	-10.80
AC	1570	0.05294	1769.3	0.05966	-11.26
AG	1566	0.05281	1908.2	0.06435	-17.93
TA	1372	0.04627	1841.5	0.06210	-25.50

*The Following Sequences Were Used, Following the Codes of Dayhoff et al. (Ref. 14):

E-GTN3 4957bp	E-GREC 1500	E-ROEC 1100	E-NNEC2 683
E-TSECA 3388	E-IS2EC 1438	E-RPECW 932	E-RDEC 667
E-RNECB 2874	E-RQECA 1320	E-RP12E 843	E-ILVEG 570
E-DEECK 2463	E-RDECD 1200	E-IS1EC 768	E-RRNC 500
E-ECOR1 2214	E-RPECL 1113	E-RRNB 720	E-RP7GE 405

Duplet asymmetries have been observed in smaller samplings on numerous occasions, most recently by Nussinov (9). There are differences between the asymmetries in E. coli and yeast sequences, yet these asymmetries are remarkably conserved within each genome. The frequencies of all 16 duplets are given in Table 1 for 20 different E.coli sequences with a combined length of almost 30kbp. The standard deviation of observed from expected frequencies is $\pm 15\%$ for the 10 duplet base pairs. This variance is six times larger than the variance for duplet frequencies in a random sequence of the same length. Several duplets have been clearly favored and others disfavored in the evolution of the E. coli genome. For example

$\begin{matrix} 5' & 3' \\ <GC> \\ 3' & 5' \end{matrix}$ and $\begin{matrix} 5' & 3' \\ <AT> \\ 3' & 5' \end{matrix}$ appear 20.2% and 16.4% more frequently than expected, while $\begin{matrix} 5' & 3' \\ <TA> \\ 3' & 5' \end{matrix}$ and $\begin{matrix} 5' & 3' \\ <GC> \\ 3' & 5' \end{matrix}$ are found 25.5% and 13.0% less frequently. Moreover, those

which are favored, disfavored and unchanged in any of the individual

sequences correlate strongly with those which are favored, disfavored and unchanged in any other sequence from *E. coli* but not with any from other species. The correlation is particularly obvious, of course, between individual sequences and the overall results for 20 sequences in Table 1. Plots of the variation in the frequency of the duplex $\begin{smallmatrix} 5' \\ <T \\ A \\ 3' \end{smallmatrix}$ with position in *E. coli* sequences shows it to be rather uniformly below the expected frequency within sequences as well.

We have been interested in establishing the relationship between the frequency of occurrence of particular oligonucleotide sequences to the frequencies of their component tuples. As expected, frequencies of long tuples show exceedingly high deviations from random expectation; which, depending on the particular evolutionary process, may or may not reflect the frequencies of smaller, component tuples. Table 2 summarizes the results of analysis for quintuplets in the *E. coli* Tn3 sequence, that exceed their expected appearance by a factor of four. Four of the twelve in this table (out of 1024 possible quintuplets), indicated by the arrow-sets in Table 2, are complementary to one another. Many quintuplets appearing in high frequency in the Tn3 sequence are found in high frequency in other sequences from *E. coli* as well. Conversely, 42 quintuplets do not appear at all in Tn3 DNA, although expected in about five copies.

We have searched all sequences for the recurrence of oligonucleotide segments of unspecified but identical sequence in different genes from the same organism as well as the same genes from different organisms. Computer

Table 2. Quintuplet Statistics for the 4957 bp Primary Strand of *E. coli* Tn3 DNA.

Quintuplet (Rank Order)	Frequency	Relative Frequency	Expected Frequency	Expected Relative Frequency	Percent Difference
TTTTT	21	0.004236	4.8	0.000968	+338
TGATA	15	0.003026	4.9	0.000983	+331
CAGCA	21	0.004236	4.9	0.000986	+330
TTTTC	20	0.004035	4.8	0.000969	+316
TCAGC	20	0.004035	4.8	0.000976	+313
TTCAG	19	0.003833	4.8	0.000975	+293
CCAGC	19	0.003833	4.8	0.000978	+292
GCCAG	18	0.003631	4.8	0.000972	+274
GCTGA	17	0.003429	4.8	0.000971	+253
CTGCC	15	0.003026	4.8	0.000963	+214
CTTTT	15	0.003026	4.8	0.000969	+212
CTGAT	15	0.003026	4.8	0.000975	+210

analysis with routine COMPARE of sequences coding for homologous proteins from different species has obvious utility in studies of particular molecular processes involved in sequence evolution. A COMPARE analysis of mouse and human hemoglobin beta chain genes for the occurrence of matching segments, for example, indicates the presence of 96 unhyphenated dodecanucleotide sequences. Most of this extensive homology is longer than 12bp, and, interestingly, cannot be made contiguous with the incorporation of hyphenated residues. We have also searched for the recurrence of identical short sequences in variant genomes of several transforming viruses, but especially between viral and host sequences. For example we have searched the 5292bp polyoma virus genome (10) and a concatenated collection of 6 sequences of more than 10kbp from the permissive mouse genome for the recurrence of identical dodecanucleotide sequences in both. The frequency of recurrence of such a long segment in two totally unrelated (random) sequences of the same base compositions as mouse and polyoma is expected to be only $1/14,880,439$ (eqn. 2). Results of routine COMPARE analysis indicate the sequence between residues 4524-4535 of the polyoma genome coding for putative coat protein VP2 is identical with the noncoding sequence between 1231-1242 of the mouse immunoglobulin kappa chain gene (11). Recognizing that extensive recombination normally occurs between polyoma and cellular DNAs (12) the recurrence of a single dodecamer is unremarkable, unless it be there is so little unhyphenated homology. However, when the search for dodecanucleotide recurrence was extended from mouse sequences to human, which is nonpermissive for polyoma virus, numerous dodecamers of identical sequence were found. Three identical dodecanucleotide sequences were found in the polyoma virus genome (10) and human hemoglobin beta chain gene (13) alone (Figure 1). There is no a priori reason to expect matching sequences of such length between these two sources, therefore, the finding of three must reflect a comparatively common path of evolution shared by these two sequences. Among the several possibilities that could account for this extraordinary homology is that a third species serves as common host to polyoma as well as for a different transforming virus that finds human cells permissive.

Palindromic sequences. Our computer search of the current DNA sequence database of more than 300,000 bp (14) indicates that palindromic sequences 10-12 bp in length with either two-fold or mirror symmetry occur ubiquitously in both coding and noncoding regions. Longer palindromes are generally rare, yet under appropriate circumstances they can be found in surprisingly high

Figure 1. Matching dodecanucleotide sequences found in both human hemoglobin beta chain gene and polyoma virus genome. The expected frequency of identical dodecanucleotides in these two DNAs is only 1/14,307,590. The beta chain gene is 2052 bp in length (13) while the polyoma genome is 5292 bp long (10).

frequency. As prescribed by eqn. 7 such circumstances occur with drifts in base composition from 50% (G+C). Thus, while perfect hexadecanucleotide palindromes are virtually nonexistent throughout most of the present database, one finds 42 in the 6241bp yeast mitochondrial cytochrome B gene sequence, which has a base composition of only 19.2%G+C (15). (Interestingly, none occur in coding regions). Of those, 21 have two-fold and 21 mirror symmetry, and many are actually longer than 16bp. A perfect docosanucleotide palindrome with mirror axis at residue 5862 has the structure

Additional minor axes occur on either side of the major axis, at 5859 and 5865, for two hexadecanucleotide palindromes with mirror symmetry. Increases in the frequency of occurrence of palindromic sequences within a DNA also coincide with significant variations in the base compositions of local domains from 50%(G+C). This variation with base composition is better appreciated with a quantitative demonstration of the variation in the expected relative frequency, p_s . A plot of equation 7 is shown in Figure 2 for the variation of the relative frequency of two-fold tetradecanucleotide ($n=7$) and octacosanucleotide ($n=14$) palindromes with base composition. After Brezinski, (8) imperfections (hyphens) are designated $\underline{h}$, where here $h=n-s$. Perfectly symmetrical tetradecamers and octacosamers are denoted 7_s and 14_s in Figure 2, respectively, while single-hyphenated tetradecamers and double-

Figure 2. Variation of the expected relative frequency of selected two-fold palindromes with base composition from eqn. 7 (line). The open circles represent the observed relative frequencies of the same palindromes in random sequences.

hyphenated octacosamers are denoted 6_{s1h} and 12_{s2h} . The open circles in this figure represent the observed relative frequency of two-fold palindromes in random sequences of 100-150 kbp from computer analysis for palindromes by routine TWOFLD. Above $1/10^4$ where the observed relative frequencies of palindromes can be accurately determined in random sequences of reasonable length, the correspondence between observed and calculated frequencies is exact, demonstrating the efficacy of our a priori calculations. The observed frequencies of $(2n)$ and $(2n+1)$ -type palindromes in random sequences were approximately equal in all cases.

A distinctive feature of equation 7 is the strong dependence of expected palindromic frequencies on base composition. The relative frequencies of octacosamers (14_s) in Figure 2, for example, represented on a logarithmic scale, increase by four orders of magnitude between 50-0% (G+C) and 50-100% (G+C). Clearly, even small variations in base composition can be expected to lead to significant changes in the relative frequency of occurrence of palindromes. The 8024 bp genome of Cauliflower Mosaic Virus (CaMV) has an

Table 3. Two-Fold Tetradeca-Palindromes in the Cauliflower Mosaic Virus Genome.

Axis	Exceptions	Axis	Exceptions
2290	0	2092-2093	1
74	1	3663	1
93-94	1	4003	1
98-99	1	4635-4636	1
258	1	4840	1
279	1	4875	1
326-327	1	5192-5193	1
675	1	5366-5367	1
736-737	1	5381-5382	1
784-785	1	5774-5775	1
834-835	1	6058	1
849	1	6157	1
976	1	6175	1
1015-1016	1	6363	1
1146	1	6496-6497	1
1497	1	6707	1
1600	1	6789-6790	1
1790	1	7583	1
2016	1		

overall base composition of 40.0%(G+C) (16), and has palindromes clustered in regions of greatest variation. The standard deviation of the variation in base composition with length in CaMV is over two times that of a random sequence of equivalent overall base composition, reminiscent of the wide deviations in nearest neighbor frequencies seen in E.coli DNA. The deviation over the first quarter of the CaMV genome is especially large, drifting significantly below the mean between positions 0-2100. In agreement with the expectation for random occurrences of palindromes, this segment contains three times the number of $6_{s_h}1_{s_h}$ palindromes than the remaining three-quarters of the CaMV genome (Table 3). Analysis of other sequences also indicate a coincidence between deviations in the relative frequency of occurrence of palindromes and deviations in base composition.

Table 3 indicates the existence of one perfect tetradeca-palindrome with center of symmetry at residue 2290 in CaMV. The expected frequency of this palindrome is $1/6214$ and $1/312$ for single-hyphenated tetradeca-palindromes for DNA with the G+C content (40.0%) of CaMV. These frequencies compare very well with observed values of $1/6682$ and $1/286$ from a TWOFLD analysis of a random sequence of 100 kbp with a G+C content of 40.0%. As seen in Table 3, the observed frequencies in CaMV DNA are $1/8024$ and $1/223$ for the perfect and single-hyphenated tetradeca-palindrome, respectively; that is, not significantly different than expected from chance occurrence.

Since approximately one such perfect palindrome is a priori expected in CaMV, we cannot assume this sequence to have any special functional or evolutionary significance; nor, of course, can we assume that it doesn't.

Analysis of CaMV DNA for tetradeca-palindromes with mirror symmetry indicates three unhyphenated and exactly the same number of $6_{s_h}1$ single-hyphenated palindromes as there are of the twofold type (Table 3). One perfect mirror palindrome with symmetry axis between residues 7621-7622 with the sequence

```

      ....ATTTCCTAAAAC|CAAAATCCAGT....
      ....TAAGGATT|TTGGTTT|TAGGTCA....
      7610          7620          7630
  
```

is actually a hexadecamer. While the number of hyphenated palindromes is not significantly different from random expectation, the occurrence of a perfect hexadecamer is noteworthy, as it is expected in a frequency of only 1/23,896. Unfortunately, the genetic map for this virus lacks sufficient detail to warrant consideration of possible functions.

Origins of Replication. There are reported examples of palindromic sequences in the vicinity of the origin of replication (O_R) in several viral genomes. Those reports have been accompanied by suggestions of unusual secondary structure in the DNA permitted by the palindromic sequence; implying some functional purpose for these structures, perhaps in the recognition of the origin or in the facilitation of the initiation process. Dhar et al (2) and Soeda et al (3) noted the existence of almost identical palindromes near the O_R loci in Bk, SV40 and polyoma viruses. In addition to homology of sequence near the O_R these three viruses are related by their ability to produce tumors, raising the question whether a palindrome at the O_R is present by accident in these viruses or whether it is a universal feature of all O_R and if so whether it can be said to play some functional role. Accordingly, we conducted an extensive computer search using routine TWOFLD to locate palindromes with two-fold symmetry in 20 sequences in the present database containing an O_R locus. With exception of just three single-stranded coliphage each of these sequences had a large palindrome either at or very close to the putative O_R locus. The palindromes with their position in the genome or genomic segment are given in Table 4 for 17 different species or strains. Also in this table are the positions of the O_R locus in the genome as reported. The fourth column gives the probability for obtaining a symmetry at a given point in the palindrome (eqn. 6), while the fifth column records the level of symmetry and the final

column the probability for the occurrence of that palindrome in a DNA of random sequence and identical base composition as the genomic specimen.

In only three occasions the palindrome appearing at the origin was surpassed in length and symmetry by another palindrome in the same or different sequence from the same species. Coli plasmids pA03(A02) and pBR322 have extraordinarily large palindromes a short distance from the O_R . Their sequences (Table 4) are almost identical; not surprising, perhaps, considering the very recent pedigree of the O_R segment in pBR322. Extensive homology also exist among the O_R palindromes from pA03(2) and pBR322 and also between the two polyoma strains and SV40 (2). Generally speaking, however, there is little homology among sequences of palindromes at the O_R .

The average palindromic length for all O_R listed is 28 ± 6 bp, while the average level of symmetry is $11_s 3_h$. With an average value of 0.2513 for p_{tf} we estimate the expected relative frequency (p_s) for the O_R palindrome to be $1/25,933$. The expectation of finding a palindrome of this average length and symmetry at the O_R in each of the 14 independently determined sequences in Table 4 is, therefore, $(1/25933)^{14}$. It seems likely that a palindrome will be a universal feature of the O_R locus.

Table 4 also includes a large 36 bp palindrome with $14_s 4_h$ symmetry found at residue 4003 in the 8024 bp sequence of the CaMV genome. This palindrome has a probability of $1/263,584$ of occurring by chance, and we would predict the O_R to occur near this locus in CaMV.

In a consideration of possible roles for two-fold palindromes at O_R we have ignored alternate secondary structures, such as the cruciform, as having any biological significance. In many cases symmetry extends into the diad axis, so that loops of 4 or more base pairs, shown to be required from the tRNA and synthetic polynucleotide work, are not possible without sacrifice of stacking potential. The lengths and levels of symmetry found in O_R palindromes are very similar to that in the E.coli lac operator sequence ($n=35$, $14_s 3_h$). A further similarity is the universal appearance of all four repeating duplets in approximately twice the expected frequency. Perhaps, replication is initiated by a tetrameric protein with C_2 symmetry, giving the protein a significantly greater bonding potential for the O_R locus.

Repeated Sequences. A search for repeated sequences is a search for matching segments since the translational repeat length cannot be a priori specified. Results of an illustrative analysis of the 470bp bovine pre-proparathyroid hormone gene (preproPTH) (17) for matching heptanucleotide sequences indicates the presence of five unique matching heptameric and two

Table 4. Characteristics of Two-fold Palindromes at Origins of Replication.

source	approximate position of O_R	palindromic sequence	p_{1f}	symmetry	relative frequency (p_s)
<i>S. typhimurium</i> (19)	127-422	...GGCTATTAA ¹²⁰ AAAGA ¹⁴⁰ TC ¹³⁰ TTT ¹⁵⁰ TTT ¹⁶⁰ AAAGAT ¹⁷⁰CCGATAAT ¹²⁰ TTCTTCT ¹³⁰ AGAAA ¹⁴⁰ AAAT ¹⁵⁰ CTCTA ¹⁶⁰ ...	.227	11 _s 4 _h	1/8,749
<i>E. coli</i> (20)	101-522	...GGAAAG ²²⁰ GATCATT ²³⁰ ACTGT ²⁴⁰ GAATGATC ²⁵⁰ GGT ²⁶⁰ GAT... ...CCTTTC ²²⁰ CTAGTAAT ²³⁰ TGACA ²⁴⁰ CTTACTAG ²⁵⁰ CCACTA ²⁶⁰ ...	.2505	10 _s 3 _h	1/8,542
coli plasmid R100 (21)	1-1112	...CAGCCTGTCT ²²⁰ CTT ²³⁰ CT ²⁴⁰ TGACAG ²⁵⁰ CAGAT ²⁶⁰GTGC ²²⁰ GACAGGA ²³⁰ CACCTGTCT ²⁴⁰ GTCTA ²⁵⁰ ...	.2493	9 _s 2 _h	1/2,885
coli plasmid CloDF13 (22)	1-705	...CTTCTGA ⁵⁶⁰ AAACCCACCT ⁵⁷⁰ CC ⁵⁸⁰ CAAGT ⁵⁹⁰ GGTTT ⁶⁰⁰ TTCTC ⁶¹⁰ TTT... ...GAACT ⁵⁶⁰ AGTT ⁵⁷⁰ GGT ⁵⁸⁰ GAAG ⁵⁹⁰ GTCC ⁶⁰⁰ ACCAAA ⁶¹⁰ AAACAAA...	.2513	12 _s 4 _h	1/27,532
coli plasmid pAO3 & pAO2 (22,23)	403-838	...ACAG ³⁴⁰ TTATCCAC ³⁵⁰ GA ³⁶⁰ CTCAGGATAA ³⁷⁰ AGGCC... ...TGTC ³⁴⁰ AAATAG ³⁵⁰ GTCT ³⁶⁰ AGTCTCCTATT ³⁷⁰ CCGG...	.2504	9 _s 2 _h	1/8,360
		...CGGCAAA ⁹⁴⁰ AAACCCACC ⁹⁵⁰ CT ⁹⁶⁰ TG ⁹⁷⁰ TA ⁹⁸⁰ CGGTGGTTT ⁹⁹⁰ TTT... ...GCCGTTT ⁹⁴⁰ GT ⁹⁵⁰ GGT ⁹⁶⁰ GGC ⁹⁷⁰ AT ⁹⁸⁰ CCACCAAAA ⁹⁹⁰ AAACAAA...		14 _s 3 _h	1/96,478
coli plasmid pBR 322 (24)	~2500	...ACG ²⁴⁴⁰ CTTATCC ²⁴⁵⁰ C ²⁴⁶⁰ GA ²⁴⁷⁰ TCAG ²⁴⁸⁰ GGATAA ²⁴⁹⁰ CGCA... ...TGC ²⁴⁴⁰ CAATAG ²⁴⁵⁰ GTCT ²⁴⁶⁰ AGTCC ²⁴⁷⁰ CCTATT ²⁴⁸⁰ GGCGT...	.2512	10 _s 2 _h	1/27,010
		...CGGCAAA ³⁰⁶⁰ AAACCCACC ³⁰⁷⁰ CT ³⁰⁸⁰ TG ³⁰⁹⁰ TA ³¹⁰⁰ CGGTGGTTT ³¹¹⁰ TTT... ...GCCGTTT ³⁰⁶⁰ GT ³⁰⁷⁰ GGT ³⁰⁸⁰ GGC ³⁰⁹⁰ AT ³¹⁰⁰ CCACCAAAA ³¹¹⁰ AAACAAA...		15 _s 2 _h	1/13,131,565
coliphage lambda (25)	1033-1233	...TATT ¹¹⁴⁰ GT ¹¹⁵⁰ CAGAG ¹¹⁶⁰ ATT ¹¹⁷⁰ CT ¹¹⁸⁰ CTCTGAC ¹¹⁹⁰ AGCC... ...ATAA ¹¹⁴⁰ ACGACTCT ¹¹⁵⁰ TAA ¹¹⁶⁰ GAC ¹¹⁷⁰ CTTA ¹¹⁸⁰ GAGACTGT ¹¹⁹⁰ CTCG...	.2448	12 _s 4 _h	1/36,430
coliphage T ₇ (26)	113-142	...AACCT ¹²⁰ TAAAG ¹³⁰ TTT ¹⁴⁰ TAAC ¹⁵⁰ TTAAGACC... ...TGAAT ¹²⁰ TC ¹³⁰ CAAA ¹⁴⁰ TTTGA ¹⁵⁰ AAATCTGG...	.2545	8 _s 1 _h	1/8,475
rat mitochondrion (27)	1-160	...GTGT ¹²⁰ ATCA ¹³⁰ CAAG ¹⁴⁰ TTAT ¹⁵⁰ GT ¹⁶⁰ TGATCATG... ...CACAT ¹²⁰ AGTGT ¹³⁰ GTCT ¹⁴⁰ CAATAC ¹⁵⁰ ACTAGTAC...	.2537	7 _s 3 _h	1/296
human mitochondrion (28)	1-185	...TGTG ⁶⁰ AAAG ⁷⁰ AGC ⁸⁰ CT ⁹⁰ GT ¹⁰⁰ CGAC... ...ACAC ⁶⁰ CTTT ⁷⁰ CCGAC ⁸⁰ AC ⁹⁰ CGCT...	.2335	5 _s 2 _h	1/117
polyoma virus (strain 3) (29)	~1	...TATT ²⁰ TAAG ³⁰ CAGGCC ⁴⁰ GGG ⁵⁰ CCC ⁶⁰ GGCTC ⁷⁰ CCCTACTC... ...ATAA ²⁰ ATTCT ³⁰ CTCCGG ⁴⁰ CCC ⁵⁰ GGG ⁶⁰ CCGAG ⁷⁰ CCAAATGAG...	.2508	14 _s 2 _h	1/3,816,755
polyoma virus (strain a2) (3)	5263-45	...TAAGCAGA ⁵²⁹⁰ GGCC ⁵³⁰⁰ GGG ⁵³¹⁰ CCC ⁵³²⁰ GGCTTACT... ...ATTCTCT ⁵²⁹⁰ CTCCGG ⁵³⁰⁰ CCC ⁵³¹⁰ GGG ⁵³²⁰ CCGAAATGA...	.2507	7 _s 2 _h	1/446
SV 40 (30)	5118-5197	...GCT ⁵¹⁸⁰ CAGAGCC ⁵¹⁹⁰ GAAG ⁵²⁰⁰ CT ⁵²¹⁰ GGCTCT ⁵²²⁰ GCAT... ...CGAG ⁵¹⁸⁰ CTCCGG ⁵¹⁹⁰ CCG ⁵²⁰⁰ CGAGCC ⁵²¹⁰ GGAGACGTA...	.2585	13 _s	1/43,517,921
SV 40 (strain 776) (31)	~1	...GCT ⁵²¹⁰ CAGAGCC ⁵²²⁰ GAAG ⁵²³⁰ CT ⁵²⁴⁰ GGCTCT ⁵²⁵⁰ GCAT... ...CGAG ⁵²¹⁰ CTCCGG ⁵²²⁰ CCG ⁵²³⁰ CGAGCC ⁵²⁴⁰ GGAGACGTA...	.2507	13 _s	1/64,612,730
papovavirus BK (strain MM) (32)	3144-3166	...GAACT ³⁰⁴⁰ TTTATC ³⁰⁵⁰ ATT ³⁰⁶⁰ TTT ³⁰⁷⁰ CAAAAA ³⁰⁸⁰ TTGCA ³⁰⁹⁰ AAAAAAT... ...CTTGA ³⁰⁴⁰ AAATAG ³⁰⁵⁰ TAAAAA ³⁰⁶⁰ AC ³⁰⁷⁰ GT ³⁰⁸⁰ TTTAA ³⁰⁹⁰ CTTTTCTTA...	.2600	12 _s 3 _h	1/56,842
papovavirus BK (Dunlop) (33)	—	...GAACT ⁵⁰⁴⁰ TTTATC ⁵⁰⁵⁰ ATT ⁵⁰⁶⁰ TTT ⁵⁰⁷⁰ CAAAAA ⁵⁰⁸⁰ TTGCA ⁵⁰⁹⁰ AAAAAAT... ...CTTGA ⁵⁰⁴⁰ AAATAG ⁵⁰⁵⁰ TAAAAA ⁵⁰⁶⁰ AC ⁵⁰⁷⁰ GT ⁵⁰⁸⁰ TTTAA ⁵⁰⁹⁰ CTTTTCTTA...	.2608	12 _s 3 _h	1/54,912
Adenovirus - Group C (2 & 5) (34,35)	1-579	...ACCT ²⁰ TATT ³⁰ TTCCATT ⁴⁰ AAAG ⁵⁰ GAATAT ⁶⁰ GTATAA... ...TGAAT ²⁰ TAAAG ³⁰ GTAA ⁴⁰ TTCCCTTAT ⁵⁰ CTATT...	.2490	9 _s 3 _h	1/2,912
Cauliflower Mosaic Virus (16)	—	...ACT ³⁹⁹⁰ TC ⁴⁰⁰⁰ GT ⁴⁰¹⁰ CTGTAT ⁴⁰²⁰ GAA ⁴⁰³⁰ CTT ⁴⁰⁴⁰ CATACAG ⁴⁰⁵⁰ TTAC ⁴⁰⁶⁰ GA... ...TGAG ³⁹⁹⁰ ACAG ⁴⁰⁰⁰ TC ⁴⁰¹⁰ GACATCT ⁴⁰²⁰ GTAA ⁴⁰³⁰ AGTAT ⁴⁰⁴⁰ GTCA ⁴⁰⁵⁰ AAAT ⁴⁰⁶⁰ GT...	.2510	14 _s 4 _h	1/263,584

octameric sequences. This is a large number for a DNA of only 470 bp and a base composition of 40.9%G+C. The expected frequency of heptamers is 1/10,912, while the two matching octamers are expected at a frequency of

Figure 3. Sequences of selected repeating octameric and heptameric segments in bovine preproparathyroid hormone gene DNA. Numbers indicate position in the published sequence (17), and underlined segments indicate segments that recur in other (nonmatching) octa/heptamers. The expected frequency of repeating heptamers is 1/15,598 and of octamers 1/62,957 in this DNA. The sequence is 470 bp in length, with a base composition of 40.9% (G+C).

only 1/41,143. There is considerable sequence homology among nearly all matching sequences in this group (Figure 3) which could be a reflection of codon preferences in translation. The coding region consists of a 345bp segment between residues 103-448 (17), however each matched oligomeric pair in preproPTH has at least one segment in a noncoding region. Therefore, these matching sequences could also have occurred through the dysfunction of an enzyme of replication or recombination. Indeed, schemes may be devised for the origin of a large segment of this gene from short tandemly repeated elements.

Inverted Repeats. Inverted repeating sequence elements are the distinguishing feature of a wide range of transposable (insertion) DNAs that have the capability of integrating into many different sites in bacterial or plasmid genomes. INVERT analysis of the transposons Tn3, Tn9 and Tn1681 DNAs, for example, all indicate the presence of long terminal repeats of 20-66bp. Results of computer analysis of a 2874bp segment of DNA from the E.coli genome coding for the β chain of RNA polymerase (18) indicate the presence of inverted tetradecanucleotide repeating elements at the two ends of the coding region. The expected frequency of such elements in a DNA of random sequence and the same base composition (53.1%G+C) as this DNA is only 1/15,429,399.

ACKNOWLEDGEMENTS

We wish to express our thanks to L. Pratt, H.E. Blake and J.D. Blake for technical assistance; to Drs. F. Vosman, B. Griffin and M. Nickerson for helpful discussions; and to Dr. M.O. Dayhoff, National Biomedical Research Foundation, Georgetown Univ., for providing occasional access

to her DNA database. Support by the National Institute of Health (GM22827) and the MAES (Project no. 08402) is gratefully acknowledged.

* To whom inquiries about programs should be addressed.

REFERENCES

1. Wells, R.D., Goodman, T.C., Hillen, W., Horn, G.T., Klein, R.D., Larson J.E., Muller, U.R., Neuendorf, S.K., Panyotatos, N., and Stirdivant, S.M. (1980) Prog. Nucleic Acid Res. Mol. Biol. **25**, 167-267.
2. Dhar, R., Lai, C.-J. and Khoury, G. (1978) Cell **13**, 345-358.
3. Soeda, E., Arrand, J.R., Smolar, N. and Griffin, B.E. (1979) Cell **17**, 357-370.
4. Hobom, G., Grosschedl, R., Lusky, M., Scherer, G., Schwarz, E., and Kossel, H. (1978) Cold Spring Harbor Symp. **43**, 165-178.
5. Grosschedl, R., and Hobom, G. (1979) Nature **277**, 621-627.
6. Hirota, Y., Yamada, M., Nishimura, A., Oka, A., Sugimoto, K., Asada, K. and Takanami, M. (1981) Prog. Nucl. Acid Res. & Mol. Biol. **26**, 33-48.
7. Jeffreys, A.J. and Harris, S. (1982) Nature **296**, 9-10.
8. Brezinski, D.P. (1975) Nature **253**, 128-130.
9. Nussinov, R. (1980) Nucl. Acids Res. **8**, 4545-4561.
10. Soeda, E., Arrand, J., Smolar, N., Walsh, J. and Griffin, B. (1980) Nature **283**, 445-453.
11. Nishioka, Y., and Leder, P. (1980) J. Biol. Chem. **255**, 3691-3694.
12. Griffin, B.E. (1980) in "DNA Tumor Viruses" (Ed. J. Tooze) 2nd Ed, Part 2, Cold Spring Harbor Laboratory, 61-123.
13. Lawn, R.M., Efstratiadis, A., O'Connell, C., and Maniatis, T. (1980) Cell **21**, 647-451.
14. Dayhoff, M.O., Schwartz, R.M., Chen, H.R., Hunt, L.T., Barker, W.C. and Orcutt, B.C. (1981) "Nucleic Acid Sequence Database", Vol. 1, Natl. Biomed. Res. Found., Wash.
15. Nobrega, F.G. and Tzagoloff, A. (1980) J. Biol. Chem. **255**, 9828-9837.
16. Franck, A., Guilley, H., Jonard, G., Richards, K. and Hirth, L. (1980) Cell **21**, 285-294.
17. 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.
18. Ovchinnikov, Y.A., Sverdlov, E.D., Lipkin, V.M., Monastyrskaya, G.S., Chertov, O.Y., Gubanov, V.V., Guryev, S.O., Modyanov, N.N., Grinkevich, V.A., Makarova, I.A., Marchenko, T.V., and Polovnikova, I.N. (1980) Bioorg. Khim. **6**, 655-655.
19. Zyskind, J.W. and Smith, D.W. (1980) Proc. Natl. Acad. Sci. USA **77**, 2460-2464.
20. Sugimoto, K., Oka, A., Sugisaki, H., Takanami, M., Nishimura, A., Yasuda, Y., and Hirota, Y. (1979) Proc. Natl. Acad. Sci. USA **76**, 575-579.
21. Rosen, J., Ohtsubo, H. and Ohtsubo, E. (1979) Mol. Gen. Genet. **171**, 287-293.
22. Stuitje, A.R., Veltkamp, E., Maat, J. and Heyneker, H.L. (1980) Nucl. Acids Res. **8**, 1459-1473.
23. Oka, A., Nomura, N., Morita, M. Sugisaki, H., Sugimoto, K. and Takanami, M. (1979) Mol. Gen. Genet. **172**, 151-159.
24. Sutcliffe, J.G. (1979) Cold Spring Harbor Symp. **43**, 77-90.
25. Scherer, G. (1978) Nucl. Acids Res. **5**, 3141-3156.
26. Saito, H., Tabor, S., Tamanoi, F. and Richardson, C.C. (1980) Proc. Natl. Acad. Sci., USA **77**, 3917-3921.

27. Koike, K., Kobayashi, M., and Sekiya, T. (1979) Cold Spring Harbor Symp. 43, 193-201.
28. Crews, S., Ojala, D., Posakony J., Nishiguchi, J. and Attardi, G. (1979) Nature 277, 192-198.
29. Deininger, P.L., Esty, A., LaPorte, P., Hsu, H. and Friedmann, T. (1980) Nucl. Acids Res. 8, 855-860.
30. Reddy, V.B., Thimmappaya, B., Dhar, R., Subramanian, K.N., Zain, B.S., Pan, J., Ghosh, P.K., Celma, M.L. and Weismann, S.M. (1978) Science 200, 494-502.
31. Fliers, W., Contreras, R., Haegman, G., Rogiers, R., Van de Voorda, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G. and Ysebaert, M. (1978) Nature 273, 113-120.
32. Yang, R.C.A. and Wu, R. (1979) Science 206, 456-462.
33. Seif, I., Khoury, G. and Dhar, R. (1979) Cell 18, 963-977.
34. Shinagawa, M., Padmanabhan, R.V. and Padmanabhan R. (1980) Gene 9, 99-114.
35. Flint, S.J. and Broker, T.R. (1980) in "DNA Tumor Viruses" (Ed. J. Tooze) 2nd Ed. Part 2, Cold Spring Harbor Laboratory, 443-546.