J Mol Evol (1987) 25:215–229

Journal of Molecular Evolution
© Springer-Verlag New York Inc. 1987

A Model for the Development of the Tandem Repeat Units in the EBV ori-P Region and a Discussion of Their Possible Function

Samuel Karlin[1] and B. Edwin Blaisdell[2]

[1] Department of Mathematics, Stanford University, Stanford, California 94305, USA
[2] Linus Pauling Institute of Science and Medicine, 440 Page Mill Road, Palo Alto, California 94306, USA

Summary. This paper presents an analysis of the repeat units of the ori-P region of the Epstein-Barr virus (EBV) genome. These repeat units are well-conserved palindromes. The pattern of these repeats, their lengths, phases, and the distribution of the relatively few substitutions are explained by a scenario that gives a reasonable course for the evolutionary development of the pattern. The scenario suggests a model for the production of an initiating 3/2 palindrome from a moderately lengthy sequence. The palindromic units are then multiplied in judicious combinations by mechanisms of unequal crossing-over events associated with some point substitutions and a few instances of slippage replication. The potential secondary structures of the two separated tandem palindromic repeat regions in ori-P are contrasted. Possible modes of binding of Epstein-Barr nuclear antigen (EBNA) 1 protein to these hairpins are discussed. A number of possibilities for the origin and development of the ori-P region in relation to viral and cellular function are considered.

Key words: Epstein-Barr virus — Origin of latent replication — Secondary structures — Binding protein—Tandem palindromic repeats

Introduction

The Epstein-Barr virus (EBV), a human herpesvirus, is associated with infectious mononucleosis and with a number of malignancies, including Burkitt's lymphoma and nasopharyngeal carcinoma (Miller 1985). It can integrate into the human genome (Henderson et al. 1983), but ordinarily resides in a latent state reproducing as a stable episome in B lymphocytes (van Santen et al. 1981). A region of the Epstein-Barr virus genome has been proposed as the origin of latent viral replication of the episomes based on replication of plasmids containing it in cultured cells infected with the virus (Yates et al. 1984, 1985; Lupton and Levine 1985; Reisman et al. 1985). This 1998-bp region, denoted ori-P, extends from positions 7315 through 9312 of the EBV genome (Baer et al. 1984) (for easier reference, we will renumber this segment 1–1998). Ori-P is 55% A+T nucleotides, which contrasts with 40% A+T content for the whole genome. Baer et al. (1984) and Reisman et al. (1985) review some of the structural features of ori-P (see Fig. 1). Region I, a 620-bp stretch in the 5′ half of the 1998-bp region, of which the 3′ third is sufficient for ori-P function, consists of 21 tandem imperfect copies of a 30-bp unit known as the "21 × 30-bp repeats." One kilobase pair downstream, region II contains four partial copies of the same repeat unit, of which at least some part has also been shown to be essential for ori-P replication function.

The stable replication and maintenance of recombinant plasmids in latently infected cells requires the EBV encoded nuclear antigen EBNA 1. Additional evidence, including nuclease protection experiments, shows that EBNA 1 binds to and interacts with the region I and II repeat units, and probably thus activates replication (Rawlins et al. 1985; Yates et al. 1985).

Offprint requests to: S. Karlin

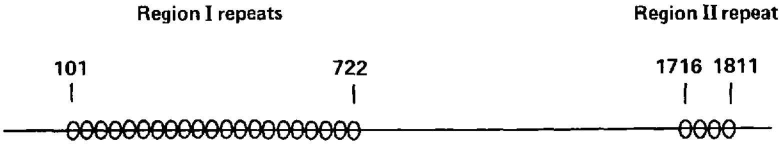

Fig. 1.

We present a detailed analysis of the repeat structures (direct and inverted) and possible secondary structures of ori-P. Although the arrangement of the 21 × 30-bp repeats roughly corresponds to tandem iteration of a basic unit of length 30 bp, the regularities and symmetries of these units, and the existence of much longer repeat units (Table 1), support a scenario of evolutionary events of sequence duplications and point substitutions that entails more than straightforward duplications of a basic unit. A plausible order of evolutionary events is proposed. We suggest that our development may well be a model for a process of DNA extension of sometime occurrence, and that in the ori-P region there is a fortuitous preservation of evidence for the development that in most cases is obscured by subsequent multiple substitutions, insertions, and deletions. The scenario also illuminates the kind and location of substitutions and may help to elucidate the influences that affect them.

Table 1. Distribution of long significant perfect repeats of the ori-P region

Length (bp)	No. of copies	First coordinate of first copy[a]	Distances (bp) between successive copies[b]
226	2	136	326
89	4	163	90, 236, 90
67	2	355	296
55	5	179	90, 60, 176, 90
40	3	235	176, 150
37	3	325	30, 296
33	7	179	90, 60, 30, 146, 90, 60
29	7	205	90, 30, 30, 176, 90, 30
25	11	179	30, 60, 30, 30, 30, 146, 30, 60, 30, 30
22	6	193	90, 60, 116, 60, 90
22	5	163	90, 176, 60, 91
19	8	193	90, 60, 30, 86, 60, 90, 60
19	7	136	60, 90, 60, 116, 60, 90

[a] Explanation of the values for the 226-bp perfect repeat. The first base of the repeat, A, occurs at location 136 in the ori-P region

[b] Distance is measured in base-pair counts between 5′ nucleotides of successive copies. Note that all the distances are multiples of 30 except one which is 4 less than this, corresponding to one truncated repeat unit of length 26 in the middle of the 21 repeat units

The ori-P region contains an impressive array of statistically significant close dyad symmetry (DS) combinations whose locations strongly correspond with the 21 × 30-bp repeats of region I and with the four truncated copies of region II (Table 2). Each of the repeat units contains a central palindrome (with sometimes a few mismatches) of length ranging from 12 to 30 bp. Since palindromes in addition to bonding within themselves can also bond with other identical palindromes, a close sequence of approximately identical palindromes can generate a plethora of alternative cloverleaf-like structures. It is postulated that some of these configurations are much more favorable to binding the EBNA 1 protein than are others (see Discussion).

In the Discussion we consider the following questions in relation to the functions of the ori-P and EBNA 1 gene regions. Why are both repeat regions required for ori-P function? What is the use of multiple repeats in this context? How prevalent is the phenomenon of tandemly repeated palindromic elements and does it serve a distinctive function? Under what conditions is the proposed model for the evolutionary course leading to the ori-P repeats a relevant process of lengthening DNA sequences?

Results

Observed Repeats

The ori-P region is rich in repeats and DS combinations. We have determined all repeats and all close and global DS pairings of length ≥9 bp (allowing for mismatches) (Tables 1 and 2).

We will use the following terminology. A specified oligonucleotide is located by the coordinate of its 5′ nucleotide. The distance between successive copies is the number of nucleotides from 5′ nucleotide to 5′ nucleotide. A distance less than the length of the oligonucleotide indicates overlapping oligonucleotides.

Table 1 lists all sets of exact repeats in ori-P of length ≥19 bp that are not completely embedded in longer repeats. Note that all the long repeats of ori-P are confined to the region extending from about 100 to 750 bp (i.e., region I repeats—the "21 × 30-bp" units). The longest exact repeat involving the region II partial copies is 15 bp long, two occurrences in region I and one in region II. The only repeats of length 30 that conserve 18 or more bases are in regions I and II.

Table 2. All exact (global) dyad symmetry pairs in ori-P of length > 10 bp and selected pairs with few mismatches[a]

Pattern	Length (bp)	Sequence	Copies[b]	Dyads[b]	Incremental distances of ordered locations[c]
M	30[d]	ATAT-TGGGTAG-ATAT-CTACCCA-ATAT	397		
L	25	ACTAACCCTAATTC-ATAGCATATG	1705	1745	
A[e]	18[d]	GGATAGCATATGCTATCC	377	673	296
A′	18[d]	GG-TAGCATATGCTA-CC	137, 197, 287, 347, 463, 523, 613		60, 90, 60, 116, 60, 90
A″	20[d]	CGGG-AGCATATGCT-CCCG	1792		
B	16	TAGCATATACTACCCA	170, 260, 496, 586	402, 698	90, 146, 90, 90, 116
B′	16	TAGCATATGCTTCCCG	1721	1792	
C	15	GGGTAGTATATGCTA	403, 699	170, 260, 436, 496, 586	90, 146, 30, 60, 90, 116
D	15[f]	GGATAGCATATGCTA	137, 197, 287, 347, 377, 463, 523, 613, 673	380, 676	60, 90, 60, 30, 86, 60, 90, 60
E	13[f]	GATAGCATATGCT	Positions listed above; 10th at 1719		60, 90, 60, 30, 86, 60, 90, 60
F	12[f]	TAGCATATGCTA	140, 200, 290, 350, 380, 466, 526, 616, 676		60, 90, 60, 30, 86, 60, 90, 60
F′	12[d]	TAGTATATACTA	1775		
G	11	GGATAGCATAT	137, 167, 197, 257, 287, 347, 377, 433, 493, 523, 583, 613, 673	384, 680	30, 30, 60, 30, 60, 30, 60, 56, 30, 60, 30, 60
H	11[f]	AGCATATGCTA	141, 201, 291, 351, 381	140, 200, 290, 350, 380, 467, 527, 617, 677, 1797	60, 90, 60, 30, 86 60, 90, 60, 466, 526, 616, 676, 1721
I[g]	11	CCCACCCCATG	1302	1666	
K[g]	10	CCCCTTGTTA	1145	1820	

[a] Mismatches are denoted by dashes. Mismatches are allowed provided there are at least five consecutive dyad matches separating mismatch positions

[b] Coordinates refer to positions of the copies or the dyads

[c] The conserved central four bases of the palindromes of regions I and II are ATAT, and these occur in all patterns A–H (underlined). The incremental distances are those between the occurrences of this fragment ordered for both the identity and dyad copies. All the increments are multiples of 30, except one that is 4 less than this, corresponding to one truncated repeat unit of length 26 in the middle of the 21 repeat units

[d] Patterns A, A′, A″, F, and F′ are palindromes. A′ and A″ have two mismatches

[e] All patterns A–H, L, and M occur in region I or II. The occurrences in 1700–1812 are in region II

[f] Pattern H is embedded in patterns E or F, which are in turn embedded in pattern D

[g] Patterns I and K are related to each other and to region II as shown in Fig. 4. I* and K* are the dyads of I and K, respectively. J_1–J_4 are the palindromic repeat regions of region II, Table 4. The gap between K and I is 147 bp and that between I* and K* is almost the same, 143 bp

Evolutionary Scenario Leading to the Repeat Structure in Fig. 1 and Tables 3 and 4

The presence of the long repeats and the periodicities in the distances separating the multiple repeat elements described in Table 1 strongly suggest that the events leading to the present repeat structure are *not* simple duplications of a canonical 30-bp unit. A near-parsimonious order of evolutionary events that could have produced the present sequence structure is proposed. Two principal types of mechanisms are utilized, tandem duplications by unequal crossing-over events (Petes 1980; Klein and Petes 1981) and point substitutions. A few local sequence iterations or deletions that may be associated with polymerase slippage and blockage are also used (e.g., Streisinger et al. 1966; Karlin and Ghandour 1985; Tautz et al. 1986).

The scenario entails a number of tandem duplications of a 30-bp unit followed by duplication events of a 90-bp segment and then of a 266-bp segment (see Table 3 for the observed sequence). The repeat units (Table 3) are all of length 30 bp, except for two of length 26. It is clear that in any lengthy tandem set of repeat units, the phase of a fundamental (original) unit that generates the tandem duplications may a priori be chosen arbitrarily. Given the repeat units as palindromic sequences, there are only two possible phases. One choice of phase is rejected, as it makes the center of the palindrome undergo many substitutions compared to the other phase that places most of the substitutions at the ends. We have fixed the phase by centering conserved exact palindromes (Table 2) in the repeat units. With this choice of phase, the 21 repeats (except for the extreme units) differ from 30-bp palindromes by at most five mispaired stem positions. The chosen phase also provides the same conserved central palindromes in the truncated repeat units of region II (Table 4).

Although it is easy to account for the occurrence of short tandem iterations by stuttering, it has proved difficult to provide reasonable mechanisms for an initial lengthy tandem duplication (e.g., Shen et al. 1981). We propose a mechanism for forming a lengthy 3/2 palindromic unit **QPQ** from any arbitrary sequence (Fig. 2), where **P** is the inverted complement (dyad) of **Q** and **PQ** and **QP** are self dyads (palindromes). **PQP** can be extended subsequently to multifold tandem (palindromic) repeat units by unequal crossings-over. Effective unequal crossing-over requires alignment and base pairing of complementary segments of substantial length. Complementarity of short segments would ordinarily not produce base pairing of sufficient stability to effect recombination. Given two or more nearby repeat segments of sufficient length, more repeat segments can be generated easily by repeated unequal crossing-over events.

On these grounds we take as our ancestral unit an exact 30-bp palindrome and propose a model that could produce this palindrome and subsequent tandem duplications of it. The 21 × 30 repeat units are not identical. Differences (base substitutions) are more numerous toward the boundaries (positions 1, 2, 29, and 30) of the 30-bp units, as would correspond with a greater likelihood of copying error near the crossover points. It is for this reason that we generally show the crossover points to be at the ends of the 30-bp repeat units, although they might occur anywhere in the aligned paired segments. Crossovers at the nonterminal points and the associated substitutions in the middle of the palindromes would produce 30-bp units that would be selected against. That cut points and ligase junctions may be more prone to substitutions is supported by experience with DNA rearrangement and translocation aberrations (e.g., as in Ig V-J joining). Because of the palindromic nature of the repeat units and the putative functional importance of hairpins formed from them, we might also expect more substitutions in positions about the junction of the loop and stem and around the base of the stem. Tandem palindromic units can dyad pair, head to tail, or whole unit with another whole unit. In the first case, for a loop of length four bases, the stem loop boundary would occur at positions 13, 14, 17, and 18, which may account for the greater number of substitutions in these positions. In the second case, the loop would occur at the ends of the 30-bp unit and would reinforce the numerous substitutions in positions 1, 2, 29, and 30 associated with the proposed crossover points.

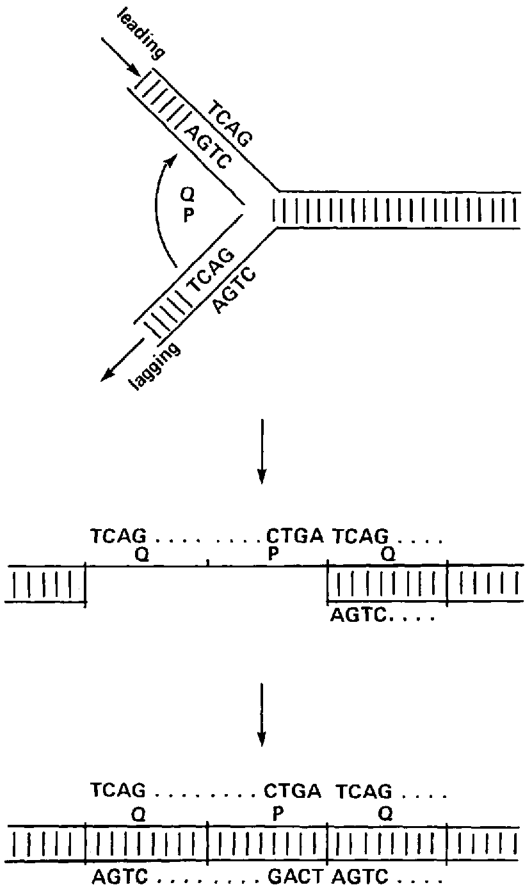
leading
TCAG
AGTC
Q
P
TCAG
AGTC
lagging
TCAG . . . . CTGA TCAG . . . .
Q
P
Q
AGTC. . . .
TCAG . . . . CTGA TCAG . . . .
Q
P
Q
AGTC . . . . GACT AGTC . . . .

Fig. 2. Mechanism of formation of initial 3/2 palindrome

Table 3 displays the sequence of region I from base 101 to base 730. Table 3 also labels each of the 21 × 30 repeat units with symbols giving a rough indication of their genealogical relationships. Occurrences of the same capital letter represent occurrences of repeat units derived from one another and differing little in sequence. Capital letters with digital superscript differ by only a single base substitution.

Initial Development of Regions I and II

Step 1

We propose as the ancestral unit the 30-bp palindrome ATATTAGGGTAGCAT ATGCTACCCTAATAT = **QP** = **W**. A model that leads to the 45-bp combination **QPQ** is shown in Fig. 2. It yields the sequence **QPQ** by following the arrows.

Table 3. 21 × 30-bp repeat sequence[a]

(coordinate 101 of ori-P)

↓

	W*	A $\underline{C}$ $\underline{T}$ $\underline{A}$ $\underline{C}$ $\underline{T}$ G G G T A $\underline{T}$ C A T A T G C T $\underline{G}$ $\underline{A}$ C $\underline{T}$ $\underline{G}$ $\underline{T}$ A T A T
	V**	$\underline{G}$ $\underline{C}$ A T $\underline{G}$ A G G A T A G C A T A T G C T A C C C G G A T A C
H	U^1	A G A T T A G G A T A G C A T A T A C T A C C C A G A T A T
	U^0	A G A T T A G G A T A G C A T A T G C T A C C C A G A T A T
	U^2	A G A T T A G G A T A G C C T A T G C T A C C C A G A T A T
H*	U^{1*}	A A A T T A G G A T A G C A T A T A C T A C C C A G A T A T
		A G A T T A G G A T A G C A T A T G C T A C C C A G A T A T
		A G A T T A G G A T A G C C T A T G C T A C C C A G A T A T
	U^0	A G A T T A G G A T A G C A T A T G C T A C C C A G A T A T
	U^{0-}	A G A T T A G G A T A G C A T A T G C T A T C C A G
	Z*	A T A T T T G G G T A G T A T A T G C T A C C C A G A T A T
	$\tilde{V}$	A A A T T A G G A T A G C A T A T A C T A C C C T A A T C T
	V*	C T A T T A G G A T A G C A T A T G C T A C C C G G A T A C
H		A G A T T A G G A T A G C A T A T A C T A C C C A G A T A T
		A G A T T A G G A T A G C A T A T G C T A C C C A G A T A T
		A G A T T A G G A T A G C C T A T G C T A C C C A G A T A T
H*		A A A T T A G G A T A G C A T A T A C T A C C C A G A T A T
		A G A T T A G G A T A G C A T A T G C T A C C C A G A T A T
		A G A T T A G G A T A G C C T A T G C T A C C C A G A T A T
	U^{0-}	A G A T T A G G A T A G C A T A T G C T A T C C A G
	Z**	A T A T T T G G G T A G T A T A T G C T A C C C A $\underline{T}$ $\underline{G}$ $\underline{G}$ $\underline{C}$ $\underline{A}$
	Q**	A $\underline{C}$ A T T A G $\underline{C}$

↑

(coordinate 722 of ori-P)

[a] The labels for the rows (30-bp repeat units) are those assigned in the detailed evolutionary development scenario described in the text

Table 4. "4 × 20-bp" repeat region II

P**	1700–1716	C$\underline{T}$T$\underline{T}$T$\underline{A}$CTA$\underline{A}$CCCTAAT
J_1	1717–1737	$\underline{TC}$G$\underline{A}$ TAGCATATGCT$\underline{T}$ CCC$\underline{G}$T
J_2	1738–1758	$\underline{T}$GGG TA$\underline{A}$CATATGCTA $\underline{TTGAA}$
G	1759–1771	TTAGGG $\underline{T}$TAG$\underline{TCT}$
J_3	1772–1792	GG$\underline{A}$ TAG$\underline{T}$ATAT$\underline{ACTA}$CTA CCC
J_4	1793–1810	GGG $\underline{A}$AGCATATG---CTA CCC
Q**	1811–1825	$\underline{GT}$TTAGGG $\underline{T}$TA$\underline{A}$CA$\underline{A}$

Legend: The first pair of repeat J units are 21 bp and the second pair 18 bp in length. The underlined nucleotides of J_1–J_4 differ from the ancestral central 18- or 20-bp palindrome. The underlined places of P**, G, and Q** indicate unmatched positions from the relevant part of P*, Q*, and Q, respectively

Table 5. Concentrations of acidic and basic amino acids near the long Gly, Ala repeat sequences, bases 286–975, in EBNA 1

Coordinates	No. of amino acids	% Acid	% Base	% Proline
1–81	27	19	4	11
82–261	60	3	27	10
262–981	240	0	0 exclusively Gly, Ala	0
982–1221	80	4	31	10
1222–1341	40	25	1	20
1342–1431	30	3	33	3
1432–1800	123	7	12	7
1801–1926	42	40	0	10

The suggested mechanism makes a copy of a segment of DNA, **Q**, then reverses and makes a copy of the segment's complementary sequence, **P**, then reverses again and copies the first segment a second time, and finally continues forward. The length of the units **Q** and **P** is determined by the length, before returning to copy the leading strand again, of the reverse complementary copy of the lagging strand, **P**. The jump from **Q** to **P** may be caused by the failure of helicase to open sufficiently the parental double helix or of helix destabilizing protein to hold the single strands sufficiently far apart. The jump from **P** to **Q** may occur when the tension in **QP** pulls it away from the increasingly distant parent strands. The proposed operations lead to the 3/2 palindromic unit **QPQ** from an arbitrary string **P**. **QPQ** can then be extended to tandem palindromic repeats of arbitrary length by repeated unequal crossing-over events. There may be other mechanisms that could produce the 3/2 palindrome, such as using inversions, nonhomologous recombination, ligation of DNA fragments, or other rereplication events. A 3/2 palindromic sequence (several kilobase pairs in length) has been found in λ-DV plasmids (Chow et al. 1974).

Step 2

QPQ is extended by an unequal crossing-over (based on the alignment of **Q**), as shown. The underline connects segments of the extended sequence, and shows the crossover at the end of a repeat unit. The crossover may occur anywhere in the dyad-paired stretch, but the chosen crossover points seem likely in view of the preponderance of substitutions near the ends of the repeat units in the subsequent development (Table 3).

$$\begin{array}{l}\underline{\mathrm{QPQ}}\\ \quad\;\mathrm{Q}\,\underline{\mathrm{PQ}}\end{array}$$

Thus we obtain **QPQPQ**, which we rename **WWQ**, **W** = **QP**.

Step 3

A further crossing over, as shown,

$$\begin{array}{l}\underline{\mathrm{WWQ}}\\ \quad\underline{\mathrm{WWQ}}\end{array}$$

yields **WWWQ**.

Step 4

By another application of the procedure of step 3, the sequence is extended to **WWWWQ**.

Step 5

We now assume a large insertion into the third **W** of about 1000 bp that separates the sequence into two parts: **S** = **WWQ*** and **T** = **P*WQ**. The insertion is made between bases 8 and 15, say, after 8, so that **Q*** = ATATTAGG and **P*** = GTAGCA TATGCTACCCTAATAT. Here we continue with the further development of S, region I. The discussion of T, region II, is given below.

Further Development of Region I

Step 6

A further **W** is generated, as shown, with nucleotide T being replaced by G at position 2 near the crossover point

$$\begin{array}{l}\underline{\mathrm{WWQ^*}}\\ \quad\underline{\mathrm{WWQ^*}}\end{array}$$

yielding $\mathbf{W\tilde{W}WQ^*}$, $\mathbf{\tilde{W}}$ = $\mathrm{A\underline{G}ATTAGGGTAGCAT}$ ATGCTACCCTAATAT, where the newly substituted base is underlined.

Step 7

A slippage deletion now occurs as described by Efstratiadis et al. (1980), using a mechanism proposed by Streisinger et al. (1966). The former authors observed many instances where a deletion in one sequence aligned with a similar sequence is bounded by short, two- to six-base, direct repeats. The deletion removes one of the repeats entirely and either none or part of the other repeat. Such a deletion occurs at the boundary between $\mathbf{\tilde{W}}$ and **W** based on the repeat ATAT, as shown,

$$\ldots\mathrm{CCCTA}\boxed{\mathrm{ATAT}}\mathrm{ATATTAGGG}\ldots$$
$$**$$

where the deleted bases are boxed. The deletion is accompanied by substitutions of A for T and of G for A at the adjacent * positions. The 5′ resulting 26-bp unit is named $\mathbf{W^-}$ = AGATTAGGGT $\mathrm{AGCATATGCTACCC\underline{AG}}$. A substitution of T for A at position 6 occurs in the **W** after $\mathbf{W^-}$ with a compensating substitution of A for T at position 25 maintaining the exact palindrome. The resulting element is named **Z** so that now $\mathbf{S = WW^-ZQ^*}$.

Step 8

The central palindrome appears to be important for the function of these repeats and is largely conserved during our development. The few substitutions in it can be argued for as follows. A viable stem loop structure would require at least a 4–6-bp loop, which, for the palindrome at hand, centers on the tetramer ATAT or hexamer CATATG. One might expect more nucleotide substitutions in the loop section, especially abutting the stem and also at the base of the stem, compared to other positions in the repeat unit. We postulate next two transition substitutions in **Z**, T replacing C at position 13, the boundary of the central hexamer, and G for A at position 26; thus **Z** → **Z***. We also assume a transition substitution in $\mathbf{W^-}$, A replacing G at bp 9, thus $\mathbf{W^-}$ → $\mathbf{\tilde{U}^-}$. [Nonterminal substitutions could also occur attendant to homologous (equal) crossing-over events.] After these substitutions we have $\mathbf{S = W\tilde{U}^-Z^*Q^*}$, where

$\mathbf{\tilde{U}^-}$ = $\mathrm{AGATTAGG\underline{A}TAGCATATGCTACCCAG}$

and

$\mathbf{Z^*}$ = $\mathrm{ATATTTGGGTAG\underline{T}ATATGCTACCCA}$ $\mathrm{\underline{G}ATAT}$.

Step 9

There is further extension by unequal crossing-over between bases 4 and 5 of **Z*** and at the end of **W**, as shown

$$\begin{array}{l}\underline{\mathrm{W\tilde{U}^-ATAT}}\ \mathrm{Z^*Q^*}\\ \quad\mathrm{W^-ATAT}\,\underline{\mathrm{\tilde{U}^-Z^*Q^*}}\end{array}$$

which with substitution of T for C at position 22 of the second $\mathbf{\tilde{U}^-}$ yields $\mathbf{WU^0U^{0-}Z^*Q^*}$, where $\mathbf{U^{0-}}$ =

$\text{AGATTAGGATAGCATATGCTA}\underline{\text{T}}\text{CCAG}$ and $\mathbf{U}^0 = \tilde{\mathbf{U}}^-\text{ATAT}$. Note that the substitution converting $\tilde{\mathbf{U}}^-$ to $\mathbf{U}^{0-}$ compensates for the substitution converting $\mathbf{W}^-$ to $\tilde{\mathbf{U}}^-$ in step 9 and reestablishes an exact 18-bp central palindrome.

Step 10

There is further extension by unequal crossing-over, as shown,

$$\begin{array}{l} \underline{\mathbf{WU^0U^{0-}\,Z^*Q^*}} \\ \quad\mathbf{W}\lfloor\underline{\mathbf{U^0U^{0-}Z^*Q^*}} \end{array}$$

which with substitution of C for T at position 30 and G for A at position 25 at and proximal to the crossover point yields $\mathbf{WVU^0U^{0-}Z^*Q^*}$, where $\mathbf{V}$ is the substituted first $\mathbf{U}^0$.

Step 11

Unit $\mathbf{U}^0$ is now amplified to three copies by two unequal crossings-over. The first $\mathbf{U}^0$ then undergoes a single substitution of A for G at position 19, giving $\mathbf{U}^1$, and the third $\mathbf{U}^0$ undergoes a single substitution of C for A at position 14, giving $\mathbf{U}^2$, and yielding overall $\mathbf{WVU^1U^0U^2U^{0-}Z^*Q^*}$. Both substitutions occur near the stem loop boundary of a hairpin formed from the central palindrome.

Step 12

We abbreviate the 90-bp unit $\mathbf{U^1U^0U^2}$ as $\mathbf{H}$. Then a tandem duplication of the 90-bp unit occurs easily by unequal crossing-over, as shown

$$\begin{array}{l} \underline{\mathbf{WV}\overbrace{\mathbf{U^1U^0U^2}}^{\mathbf{H}}}\,\mathbf{U^{0-}Z^*Q^*} \\ \qquad\quad\mathbf{WV}\lfloor\underline{\underbrace{\mathbf{U^1U^0U^2}}_{\mathbf{H}}\mathbf{U^{0-}Z^*Q^*}} \end{array}$$

which with a single substitution of A for G at the crossover point in the second position of $\mathbf{H}$ yields $\mathbf{WVHH^*U^{0-}Z^*Q^*}$. The 180-bp sequence $\mathbf{HH^*}$ can conceivably be achieved through multiple duplications on a 30-bp unit modified by mutations. However, the regular distribution of the base substitutions makes this order of events unlikely. A more probable course of events entails substitutions first in one of the $\mathbf{H}$ units (among the $\mathbf{U}$ replicas as set forth in step 10) and then this entire element is tandemly duplicated, retaining the identically distributed base changes.

Step 13

A tandem duplication of a 266-bp unit (266 is 4 less than 9 × 30), as shown

$$\begin{array}{l} \underline{\mathbf{WVHH^*U^{0-}Z^*Q^*}} \\ \qquad\qquad\quad\mathbf{W}\lfloor\underline{\mathbf{VHH^*U^{0-}Z^*Q^*}} \end{array}$$

then yields $\mathbf{WVHH^*U^{0-}Z^*VHH^*U^{0-}Z^*Q^*}$. We invoke the argument in favor of this large duplication vs a process of multiple smaller duplications, for the same reasons as given in step 12.

Step 14

Next, a tandem duplication of the central $\mathbf{V}$ unit occurs by unequal crossing-over, as shown

$$\begin{array}{l} \underline{\mathbf{WVHH^*U^{0-}Z^*V\,HH^*U^{0-}Z^*Q^*}} \\ \quad\mathbf{WVHH^*U^{0-}Z^*}\lfloor\underline{\mathbf{VHH^*U^{0-}Z^*Q^*}} \end{array}$$

This crossover is attended by several clustered substitutions near the crossover point in both $\mathbf{V}$ units. At some later stage there are two transitions G to A at position 18 at the junction of the loop in the palindrome and at position 2 at the base of the stem, both in the first $\mathbf{V}$. This process yields $\mathbf{WVHH^*U^{0-}Z^*\tilde{V}V^*HH^*U^{0-}Z^*Q^*}$.

Step 15

Finally the central region $\mathbf{U^{0-}Z^*}$ align and cross-over 4 bp into $\mathbf{Z^*}$, as shown

$$\begin{array}{l} \ldots\,\underline{\overline{\mathbf{U^{1*}U^0U^2}}^{\mathbf{H^*}}\,\mathbf{U^{0-}}\,\text{ATAT}}\,{}^-\mathbf{Z^*\tilde{V}V^*}\ldots \\ \ldots\quad\underbrace{\mathbf{U^{1*}U^0U^2}}_{\mathbf{H^*}}\qquad\qquad\;\lfloor\underline{\mathbf{U^{0-}Z^*\tilde{V}V^*}}\ldots \end{array}$$

which yields $\mathbf{WVHH^*U^0U^{0-}Z^*\tilde{V}V^*HH^*U^{0-}Z^*Q^*}$.

Step 16

At various stages in the development we postulate further substitution occurrences in the array of 21 repeats achieved at the conclusion of step 15. These additional substitutions occur in the extreme repeat units 1, 2, 21, and $\mathbf{Q^*}$, and are shown underlined in Table 3. A group of 3 clustered substitutions occurs at the 5′ end of $\mathbf{V}$, and 5 at the 3′ end of $\mathbf{Z^*}$. There are also several changes at both ends of the starting $\mathbf{W}$, but these preserve a short central palindrome. These substitutions could have arisen gradually, in part by homologous crossing-over events during the long time since the early formation of these repeat units.

The nine central exact palindromes F (Table 2) of length 12 are found in repeat units $\mathbf{V^{**}}$, $\mathbf{U^0}$, $\mathbf{U^{0-}}$, and $\mathbf{V^*}$, those in $\mathbf{U^{0-}}$ being of length 18. The observed 30-bp repeat units $\mathbf{U^0}$ (five copies), $\mathbf{Z^*}$, and $\tilde{\mathbf{V}}$ differ least from the assumed 30-bp palindromic ancestor (4-bp differences).

Most of the substitutions occur at the ends of the 30-bp repeat unit in positions 1–6 and 25–30, 27 substitutions in 12 sites. Positions 7–12 and 19–24,

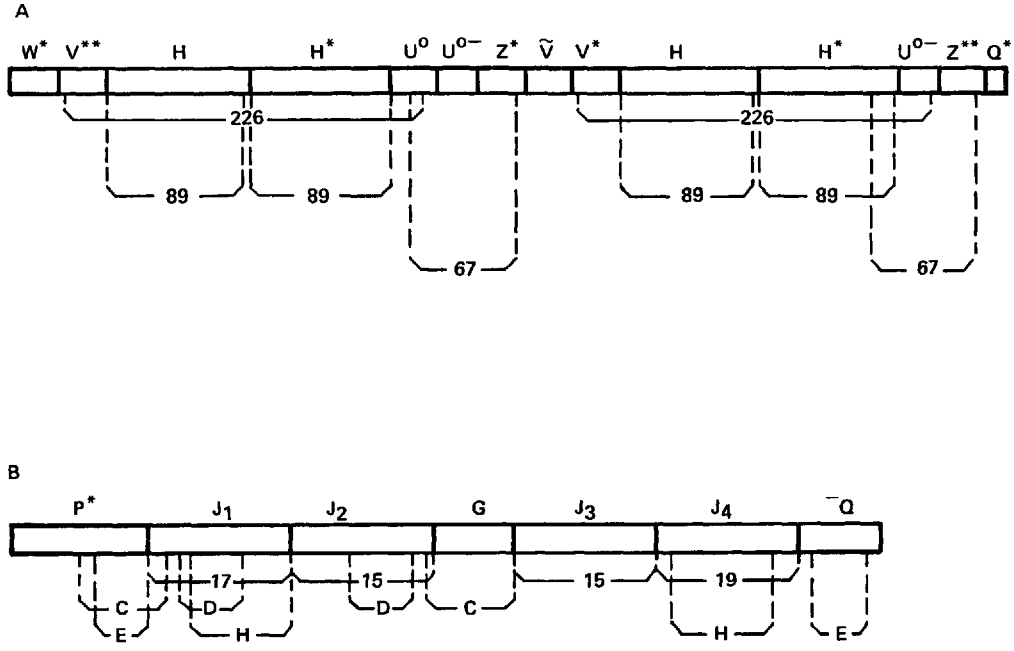

Fig. 3. Maps of regions I and II (**A** and **B**), respectively. **A** Lengths of longest exact repeats in product of scenario. **B** Row 1. Number of matches with consensus in repeat units in product of scenario (Table 4). Row 2. Exact dyads of lengths C = 14 and D = 10 from hairpin (Fig. 4A). Row 3. Exact dyads of lengths E = 9 and H = 16 from hairpin (Fig. 4B)

which might constitute the basic stem of an internal hairpin from the central palindrome, have experienced only three substitutions in 12 sites. Positions 13–18, which could form the loop of the hairpin, have experienced three substitutions in six sites. Note that those three are at positions 13, 14, and 18, at the conceivably strained internal stem loop junction. In general, there is a decline in the number of substitutions with decreasing time since formation. A notable exception is $\tilde{V}$, but four of the six substitutions in forming $\tilde{V}$ and both in forming V* are adjacent to the crossover point in their formation from U^0. A possible reason for the relatively large number of substitutions in $\tilde{V}$, V*, W, and Z* relative to U^0, U^1, and U^2 is that the long 90-bp repeat of the approximately palindromic units $U^1U^0U^2$, $U^{1*}U^0U^2$ that is capable of forming a base-paired stem with 63 paired bases out of 90 per strand may protect the $U^1U^0U^2$ units from substitution. The substitution of A for G in position 2 of U^{1*} would occur near the stem loop junction of such a hairpin. The differences between V* and V** are in the 3′ end, whereas substitutions between Z* and Z** derived from Z are in the 5′ end. The remnant fragments of Q 3′ to regions I and II are very similar to the ancestral Q, especially close to the ends of regions I and II (Tables 3 and 4). Similarly the present fragment of **P** 5′ to region II is very similar to the ancestral P, particularly close to the beginning of region II.

The relationship of Table 3 to the three longest repeats of Table 1 is shown in Fig. 3.

Further Development of Repeat Region II

The sequence of the region II repeats is considerably less regular than that of the region I repeats. Accordingly, the scenario of the evolutionary course leading to region II is less definite. Nevertheless, modulo minor variations in the developments presented below, the steps are coherent and economical.

Step 5′

As in step 5, slippage deletions in T = **P*WQ** occur at the interface between **P*** and **W**, based on repeated TA,

$$\ldots\mathrm{CCCTAATAT}\boxed{\mathrm{ATATTA}}\mathrm{GGG}\ldots$$

forming $\mathbf{P}^{*}{}^{-}\mathbf{W}$, with the boxed elements removed. Similar slippage at the boundary between ${}^{-}\mathbf{W}$ and **Q** yields the sequence $\mathbf{P}^{*}{}^{-}\mathbf{W}{}^{-}\mathbf{Q}$.

Step 6′

The sequence is then extended by an unequal crossing-over event based on the alignment as shown,

$$\begin{array}{l}\underline{\mathbf{P}^{*}{}^{-}\mathbf{W}{}^{-}\mathbf{Q}}\\ \qquad\mathbf{P}^{*}\boxed{{}^{-}\mathbf{W}{}^{-}\mathbf{Q}}\end{array}$$

yielding $\mathbf{P}^{*}{}^{-}\mathbf{W}{}^{-}\mathbf{W}{}^{-}\mathbf{Q}$, which can also be read as $\mathbf{P}^{*}{}^{-}\mathbf{W}{}^{-}\mathbf{W}^{-}\mathbf{Q}$, where ${}^{-}\mathbf{W}$ results from the contraction of **W** deleting its first six bases, and ${}^{-}\mathbf{W}^{-}$ designates the further contraction deleting seven bases from the interface of ${}^{-}\mathbf{W}$ and **Q** (cf. step 5′).

Step 7′

Another pair of crossover events subsequently expands this sequence to $\mathbf{P^{*}\,{}^{-}W\,{}^{-}W\,{}^{-}W\,{}^{-}W\,{}^{-}W\,{}^{-}Q}$.

Step 8′

The third $\mathbf{{}^{-}W}$ of the foregoing sequence undergoes a slippage event like those in steps 5 and 5′, but more lengthy, based on the repeat of AT in the center, as shown,

$$\ldots\text{GGGTAGC}\boxed{\text{ATATGCTACCCTAATTAT}}\text{TAGGG}\ldots.$$

The resulting sequence can be read (see the comment at the close of step 6′) as $\mathbf{\hat{T}} = \mathbf{P^{*}\,{}^{-}W\,{}^{-}W^{-}\,\tilde{Q}^{-}\,{}^{-}W\,{}^{-}W^{-}\,Q}$ where $\mathbf{\tilde{Q}^{-}}$ is very similar to the initial 13 bp of $\mathbf{Q}$. Some further slippage occurrences (as in step 5′) convert $\mathbf{\hat{T}}$ to $\mathbf{\tilde{T}}$:

$$\mathbf{\tilde{T} = P^{*}\,{}^{-}W^{-}\,{}^{-}W^{-}\,\tilde{Q}^{-}\,{}^{-}W^{-}\,{}^{-}W^{-}\,Q}.$$

A few local adjustments in the above sequence made to improve agreement with the observed region II can occur through point substitutions and DNA stuttering events (e.g., doubling of an existing base or doubling the triplet CTA in the third $\mathbf{{}^{-}W^{-}}$ element), yielding the region II repeats, which we represent as $\mathbf{P^{**}\ J_1\ J_2\ G\ J_3\ J_4\ Q^{**}}$ (see Table 4).

An alternative way of generating the gap segment G that avoids the more lengthy deletion of step 7′ goes as follows. The sequence $\mathbf{T}$ is extended to $\mathbf{P^{*}\,{}^{-}W\,{}^{-}W^{-}\,Q = P^{*}\,{}^{-}W\,{}^{-}W\,{}^{-}Q}$ as before. Then, based on the alignment

$$\begin{array}{l}\underline{\mathbf{P^{*}\,{}^{-}W\,{}^{-}W\,{}^{-}Q}}\\ \qquad\qquad\mathbf{P^{*}}\,\underline{|\mathbf{{}^{-}W\,{}^{-}W\,{}^{-}Q}}\end{array}$$

we postulate a crossover after bp 13 of $\mathbf{{}^{-}Q}$, but along with this recombination event the last 9-bp segment up to the crossover point of $\mathbf{{}^{-}Q}$ is duplicated (a sort of back slippage). These operations yield the sequence $\mathbf{P^{**}\,{}^{-}W\,{}^{-}W^{-}\,\tilde{Q}^{-}\,{}^{-}W^{-}\,{}^{-}W^{-}\,Q^{**}}$. Next a slippage deletion as in step 5′ converts the first $\mathbf{{}^{-}W}$ to $\mathbf{{}^{-}W^{-}}$. The remaining adjustments needed to produce the observed region II repeats paraphrase those described in step 8′.

Several of the substitutions have contributed to the strength of the stems in the hairpin structures shown in Fig. 4. In the time since their formation in steps 5′–8′, the slipped sequences occurring in positions 1705 through 1822 have undergone about 30 point substitutions, about one-third of which add pairings to the stem of the short hairpin in Fig. 4a and to the stem of the long hairpin in Fig. 4b. The hairpin in Fig. 4a, also given in Reisman et al. (1985), has 28 dyad pairs in a stem of 31 bases and a very small loop of 3 bases. This hairpin pairs the first $\mathbf{J_1}$ with $\mathbf{J_2}$ and the leading $\mathbf{P^{*}}$ with the gap G. The longer hairpin in Fig. 4b has 39 dyad pairs in a stem of 51 bases and a more feasible loop of 13 bases. This hairpin mainly pairs the first doublet $\mathbf{J_1 + J_2}$ with $\mathbf{J_3 + J_4}$, leaving G to form a generous loop. A possible function for this longer loop is discussed below. Only 5 of the approximately 30 substitutions occur in the central 12 base palindromes of ancestral repeat units $\mathbf{{}^{-}W^{-}}$. The consensus central palindrome of these four observed repeat units is TAGCA TATGCTA, which agrees with the nine central exact palindromes F (Table 2) found in region I.

Analysis and Discussion

For the ori-P region I (the "21 × 30-bp repeats") and region II ("4 × 21-bp repeats") we have been able to provide a reasonable, simple, economical, and possible course of development using only a few mechanisms that can coherently explain the observed facts about the tandem repeat units and the point substitutions within them. We suggest that the detailed scenario proposed may afford a model for a process of DNA lengthening that is of frequent occurrence. The mechanisms of DNA lengthening are ordinarily obscured by multifold point substitutions, deletions, and insertions. Only a lucky chance has caught the process in ori-P at a stage where evidence for the development can reasonably be deciphered. The scenario also illuminates the kind and location of substitutions and the influences that affect them. Our proposed scenario relies mostly on tandem duplications by unequal crossings-over and on point substitutions. Introduction of deletions and insertions by a polymerase slippage event between nearby short repeats is invoked on a few occasions early in the scenario.

The major events of the DNA extensions leading to the region I repeats include a number of tandem duplications of the "ancestral" unit ATAT TAGGGTAGCATATGCTACCCTAATAT = $\mathbf{QP}$ (this is a perfect 30-bp palindrome) and subsequent tandem duplications of a 90-bp element and of a 266-bp segment. The remnants closest to the ancestral unit in the observed ori-P sequence are $\mathbf{U^{0}}$ (five copies), $\mathbf{Z^{*}}$, and $\mathbf{\tilde{V}}$, each with two mispaired positions in the hairpin form. All repeats of region I are tandemly arranged 30-bp units except for two truncated to 26 bp.

The region II repeats maintain the core of the 30-bp palindromic units in two pairs of about 20-bp tandem copies separated by a 13-bp gap. The gap oligonucleotide is substantially similar to the ancestral unit $\mathbf{Q}$. Abutting 5′ to the first 20-bp copy is a 17-bp oligonucleotide, the 3′ thirteen of which are similar to the oligonucleotide $\mathbf{P}$, i.e., almost the same as the inverted complement of the gap.

The most highly conserved part of the 30-bp re-

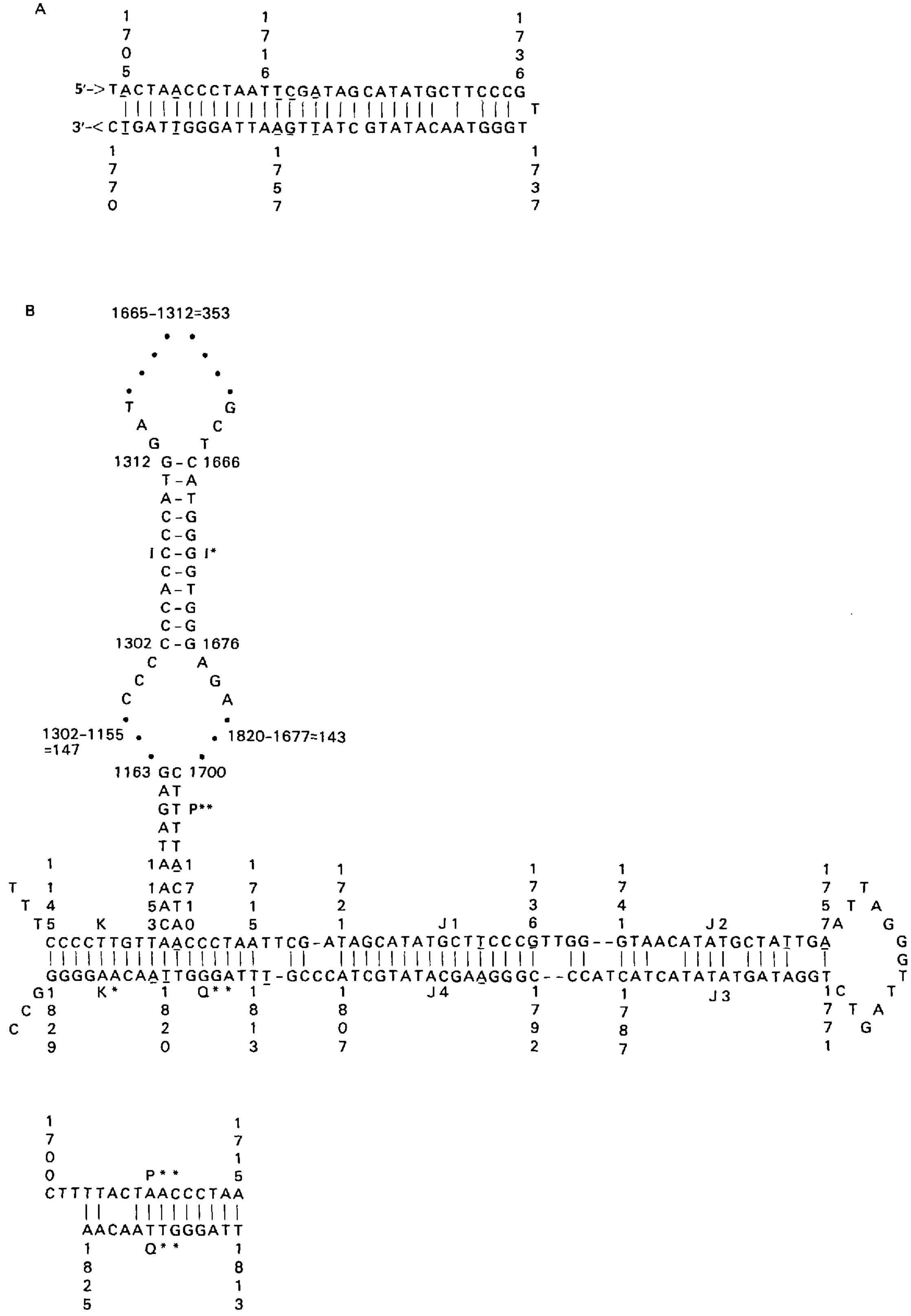

Fig. 4. Long dyad pairings in region II of ori-P. **A** Rawlins et al. (1985). **B** Our proposal. Segments are labeled with the names used in Tables 2 and 4. Note that the right end of K, TTGTTA can be paired with the right end of K* or with the left end of Q**, as shown, both of which read AACAAT. **C** Alternative pairing of the right end of P** with the right end of Q**

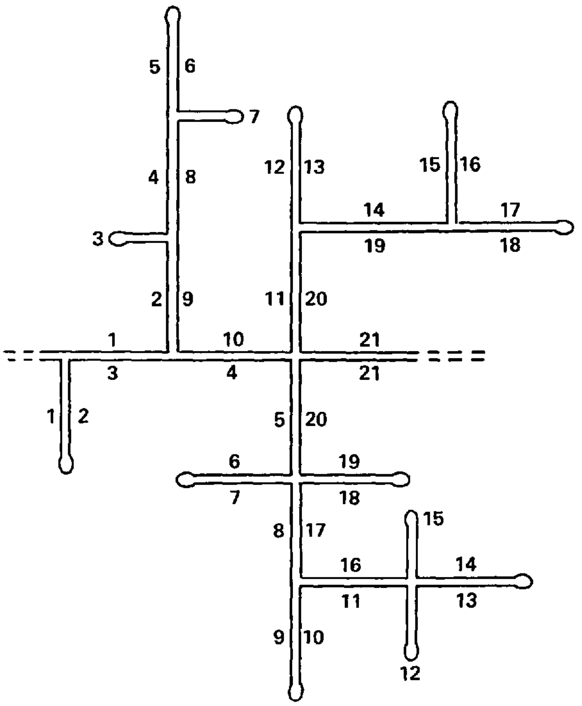

Fig. 5. An example of a possible cloverleaf structure formed from 21 palindromic repeats. The 21 tandem palindromic 30-bp repeats are numbered 1 to 21 sequentially. The cloverleaf from one strand is above the dotted nonrepeat extensions; from the other strand it is below. The half-length double helices are formed internally within one palindrome; the unit-length double helices are formed from two palindromes

peat units in both regions I and II is the 18-bp center of the unit. This center is approximately palindromic in all repeat units. An exact 12-bp palindrome occurs in 11 of these, 9 in region I and 2 in region II (Table 2). If one or two mismatches are permitted, still longer approximate palindromes are found. Repeat units 2–20 of region I differ from an exact palindrome by at most five mismatches per folded 15-bp half, mostly in positions 1–6 (and 25–30). The region I palindromic repeats have the potential for a wide assortment of extensive secondary structures (Fig. 5). The arrangement of region II possesses mainly two strong dyad symmetry combinations, described below, which may have functional value (Fig. 4). Note that the distances between **I** and **K** and between **I*** and **K*** are nearly the same, 147 and 143, respectively, thus the pairs **I**,**I*** and **K**,**K*** make an approximately aligned double dyad pairing embracing the long dyad pairing of $\mathbf{J}_1$ and $\mathbf{J}_2$ with $\mathbf{J}_4$ and $\mathbf{J}_3$, respectively (Fig. 4B). The dyads **I** and **K** are longer than any other dyads in the 2000-bp ori-P region, except those in regions I and II.

Most of the substitutions between repeats occur near the assumed crossover points, which are positions of strain and more liable to replication error. Each repeat unit has a palindromic structure that permits the formation of head-to-tail hairpins in each DNA strand. In such a hairpin there will be a region of strain at the base of the stem, corresponding to the crossover points and supporting more frequent substitutions in these regions. There will also be a region of strain at the junction of the stem and the short minimal central loop. The loop is assumed to be of length about 4 bp, since a shorter loop in the DNA strand is unstable and a longer loop unnecessarily loses base pairs in the stem. Some substitutions are found near this possible junction. The few substitutions in the most highly conserved region of the stem of a possible internal hairpin structure, positions 7–13 and 18–24, are mostly transitions. The substitutions in the less conserved region at or near the boundary between repeat units, positions 1–6 and 25–30, are more numerous and about equally transitions and transversions.

Preferences in the distribution of substitutions may be investigated by comparison with the ori-P of other strains of EBV. Rawlins et al. (1985) studied the ori-P region in two stocks of EBV: one, pHEBO-1, obtained from Sugden's laboratories (Yates et al. 1985), and a second, a clone, pPL7, of their own construction. Both were derived originally from the B95-8 strain. When the region I repeats of the two clones are aligned at an NcoI cut at their 3′ ends, the seven complete 30-bp repeat units reported for the pPL7 clone differ in 13 instances in end sites 1–6 and 25–30, but in only 3 instances in internal hairpin stem sites 7–12 and 19–24 and in 5 instances in hairpin loop sites 13–18, a pattern like that observed during the development described above. These latter counts accept the suggestion of Rawlins et al. (1985) that repeat units −5 to −2 from the NcoI cut are inverted complements of the canonical repeat. It is very surprising that there are no differences in the 126 bases of regions II compared with about 10% difference in the region I repeats. This difference in conservation is compatible with the suggested different functions of regions I and II discussed below.

Unequal crossing-over and gene conversion are two mechanisms commonly advanced to explain homogeneity between DNA segments. One might expect more perfect copy units under gene conversion than is observed in the 21 × 30 repeats. It is conceivable that the 266-bp repeat, which arose in our scenario by a relatively recent recombination event, could also have resulted through gene conversion. However, the exact 30-bp lengths of the units argue more for an unequal crossing-over mechanism.

The multiplication of an initial two-unit tandem repeat of considerable length to more units by unequal crossing-over is familiar and has been documented in other studies (e.g., Petes 1980; Klein and Petes 1981), but it is difficult to think of a mechanism for producing the initial two tandem repeat units (cf. Shen et al. 1981). However, the repeats in ori-P have a palindromic structure that enables a reasonable mechanism for producing an initial 3/2 palindromic unit **PQP** from a moderately lengthy sequence (consult Fig. 2 and step 1 of the scenario).

The scenario postulates the evolution of both the region I and region II repeats from a single ancestral unit, which, after some tandem duplications, was separated (through an unknown mechanism) by an insertion element of about 1 kbp length. The two repeat regions subsequently evolved as in steps 5–16 of the scenario for repeat region I and steps 5′–8′ for repeat region II. A tentative hypothesis alternative to separation in the virus proposes that the conglomerate of regions I and II with the intervening region is a relict of an ancestral transposon contributed by the human host or some other autonomous replicating source. Transposons characteristically show moderate-length inverted repeats at their flanks, while the recipient host DNA at the entry site of a transposon generally acquires short flanking repeats. The palindrome repeat units of region I and region II are obviously consistent with these patterns.

On the basis of the hypothesis of a transposon precursor element initiating the ori-P region, we might suppose that EBV formerly existed in a lytic productive cycle. But, when ori-P came in from the host, thus providing EBV with a natural autonomous origin of replication, a concatenation of selective forces converted EBV into its latent growth existence. The suggested relationship between ori-P and the host raises a number of interesting questions. Can ori-P-like structures function as origins of replication in the eukaryotic host? If ori-P is derived from the host, and since it requires EBNA 1 for its function, perhaps proteins like EBNA 1 are used in host replication and DNA regions like the ori-P and EBNA 1 genes can be sought by hybridization experiments. Such hybridization experiments have been carried out by Heller et al. (1985). Hybrids of the 700-bp glycine–alanine repeat region of the EBNA 1 gene have been found with several distinct regions in the human and mouse genome. When these regions were sequenced, the only substantial similarity to the EBNA 1 gene was to the 700-bp tandem repeats exclusively of glycine and alanine in EBNA 1. It is apparent that the hybridization was dominated by the presence of uninterrupted regions containing high concentrations of GG and GC doublets. However, none of the human or mouse regions are similar in higher-order structure to the EBNA 1 region. Because of the simple sequence character of the 700-bp glycine–alanine repeat region that is, in any case, not essential to its latent plasmid replication activity (Yates et al. 1985), it might be better to seek genomic hybrids under less stringent conditions with EBNA 1 DNA from which the long neutral glycine–alanine stretch has been removed. Reisman and Sugden (1986) have suggested that the induction of fgr proto oncogene mRNA in B-lymphocytes infected with EBV (Cheah et al. 1986) may be caused by EBNA 1. Also, antibodies to EBNA 1 protein have been found to react with a cellular protein of rather dissimilar peptide composition and conversely (Luka et al. 1984).

Does the existence of many eukaryotic origins of replication indicate a relation to transposons? Suggestions of this kind have been ascribed to Alu interspersed sequences. Are multiple sets of close palindromic repeats important for origins of replication? How generally do temperate episomes possess structures like ori-P? We are not aware of any hybridization experiments searching for similarity to the ori-P region in the human genome or general mammalian viruses. The present discussion suggests that such a search might be informative. Rawlins et al. (1985) made a computer search of the entire EBV genome and of the human DNA sequences in GenBank for regions similar to the central 20-bp palindromic repeat unit and found two approximate tandem copies at about coordinate 64,000 bp in the EBV genome. However, we caution that the length of the human sequence currently available in GenBank is so short (about 10^6 bp compared with an estimated total of 3×10^9 bp) as to make it unlikely to find similarities with the ori-P region that will pass a reasonable significance test. Parenthetically, we did compare the ori-P of EBV with the putative origin of replication in the long arm of herpes simplex virus type I (Gray and Kaerner 1984), but no special similarity was revealed (data not shown).

Reisman et al. (1985) and Yates et al. (1984) have found that deletion of all of region II or deletion of all of region I is each sufficient to prevent latent replication of EBV ori-P-containing plasmids in the presence of EBNA 1 protein. This protein binds to the central palindrome [protects it from DNase digestion (Rawlins et al. 1985)]. However, deletion in region I leaving only the six 3′ repeat units does not inhibit latent replication (Reisman et al. 1985). Do these facts imply that the recent exact duplication of 266 bp (almost half of region I), step 13 of the scenario, primarily provides insurance to guarantee successful latent replication? There is a phenomenon analogous to the 266-bp exact duplication in ori-P in the 72-bp exact duplication of the enhancer element of SV-40, proximal 5′ to the origin of replication. This sequence displays two perfect tandem copies, albeit only one unit is required for efficient SV-40 replication. The counterpart of EBNA 1 binding is that of the T-antigen protein that binds to the roughly 100-bp-long origin of replication of SV-40 at several places (Watson et al. 1983), and thus promotes its replication. The SV-40 origin of replication also contains a long palindrome of 26 bases, with one odd base at the center. Another possible analogue to the ori-P–EBNA 1 interaction in EBV to effect episome maintenance of the EBV genome is the episome maintenance of the bovine papilloma virus (BPV) genome. In the latter case, a 500-bp region containing an origin of episome replication has been identified (Waldeck et al. 1984), and a companion region within the open reading frame E1 has been found to be essential for episome maintenance (Lusky and Botchan 1984).

It would be interesting to determine if the 5′ half of region I is sufficient to permit efficient latent replication, as is the 3′ half, and to determine how much more region I can be shortened and still maintain proper replication. Are any four repeat units sufficient? Can any part of region II be deleted without inhibiting replication? All but the 3′ 28 bp of the 983-bp intervening sequence between regions I and II was deleted without preventing growth of host cells infected with plasmids containing selectable markers and the deleted ori-P region. However, the data indicated that the number of plasmid EBV copies was substantially diminished (Reisman et al. 1985). These results suggest that the 1000-bp insertion may play a role in determining the number of plasmids per cell, but not in cell replication. In the analogous BPV system, a region in open reading frame E7 immediately 5′ to E1 has been found to influence episome copy number. Changing the content, orientation, and size of the sequence between the region I and region II repeats may help elucidate its role.

Does the occurrence of multiple repeat units serve a useful function other than that of insurance redundancy? This may be a feature compensating for the small number of EBV episomes in the cell during latency. Although EBNA 1 protein binds strongly to the central 20 bp of the 30-bp repeat units, it also binds weakly and nonspecifically to DNA generally (Rawlins et al. 1985), possibly because of the substantial positively charged regions within it (Table 5). Multiple repeat units permit a cooperative interaction between bound protein molecules that produces a group binding much stronger than the sum of single bindings, thus aiding the virus in its competition with the entire host genome for EBNA 1. Cooperative interaction between EBNA 1 molecules is reasonable, since they contain separate substantial regions of both positively and negatively charged residues (Table 5). In the analogous BPV system, the 732-residue protein corresponding to the open reading frame E1 has a similar distribution of charge: a stretch of 35 residues near the beginning and one of 45 near the end are very strongly negatively charged. The 489 residue segment between these stretches appears to favor positive charge: 77 positive to 44 negative but without significant concentrations anywhere.

Does the palindromic nature of the repeat units serve a useful function? The origins of DNA replication in many DNA viruses (e.g., SV-40, polyoma, and herpes simplex virus types, all lytic, and the ori-P of EBV, latent) contain several palindromic and close dyad pairings. In the case of single-strand-DNA adeno-associated viruses, it has been suggested that the secondary structure inherent in the origin of replication can lead to a hairpin configuration that will provide DNA polymerase with the required primer for the initiation of replication (Hauswirth and Berns 1979; Fraenkel-Conrat and Kimball 1982).

The 30-bp repeat unit region I of EBV is high in weak-bonding A+T content (63%) compared with all other regions in EBV with an overall content of 40% in A+T. This means that the increase in energy associated with reconfiguring the DNA double helix into cloverleaf structures is not large. Furthermore, since palindromes, in addition to bonding first half to second half within themselves, can also bond whole to whole with other identical palindromes, a tandem sequence of 21 approximately identical palindromes can form a very large number of alternative cloverleaf-like structures. An example is sketched in Fig. 5. It is possible that some of these configurations are much more favorable than is the original linear DNA duplex to the cooperative binding of many EBNA 1 molecules, each bound to the center of a 30-bp hairpin stem of composition nearly the same as that of the respective linear DNA duplex repeat units. Such cooperative EBNA 1–EBNA 1 interactions may occur between some of the three positive regions on one molecule with some of the three negative regions on others. The occurrence of three complementary pairs of charged regions on each EBNA 1 protein molecule (Table 5) permits the formation of complicated three-dimensional (cooperative) structures. The binding of EBNA 1 to synthetic monomeric, dimeric, and trimeric oligonucleotides agreeing with the consensus repeat unit U^0 at bases 5–24 has been found to be cooperative (Milman 1986). The decrease in energy consequent upon interaction among all these favorable bindings, together with the entropy increase consequent upon the many configurations available, may well form stable structures able to compete with the entire host genome for the EBNA 1 molecules essential for latent replication. This structure contrasts with that possible with the binding of large T-antigen to the origin of replication in SV-40. The 72-bp repeats of SV 40 and the 21 × 30-bp repeats of EBV are alike in being effective enhancers of the production of chloramphenicol acetyl transferase (CAT) in plasmids containing both the enhancer, the gene, and a suitable promoter, but the EBV region is effective only in the presence of EBNA 1 protein (Reisman and Sugden 1986). The region near coordinate 64,000 bp in the EBV genome found to be very similar to the central 20-bp palindrome, and like it also containing a TAG terminator, may also serve as an enhancer for nearby genes (Rawlins et al. 1985). These authors further observed that the 72-bp enhancer from SV 40, which is not a global palindromic element, cannot substitute for the enhancer ori-P region I from EBV for the replication of plasmids containing region II, even in the presence of EBNA 1. This result supports the notion that the palindromic nature of region I is important for plasmid replication.

Does the existence of the two repeat regions I and II 1 kbp apart serve any useful function? As mentioned previously, neither can be completely deleted without losing the ability for latent growth. A single EBNA 1 binding at two separate sites surely entails greater binding strength, as does holding with two hands vs one hand. But this may not be the primary function of having two repeat regions. Because the palindrome repeats of region I are contiguous, we would expect that all secondary structure formations of this region are tightly knit (i.e., are almost entirely base paired with minimal loops; cf. Fig. 5), and covered with bound EBNA 1 proteins so that initiation of RNA primer is inhibited. EBNA 1 as a protein binding to double-stranded DNA surely enhances the protection of this region. On the other hand, the region II repeats merely consist of two 21-bp pairs of palindromic tandem repeats separated by a 13-bp gap. Thus, region II can form a hairpin loop based on a strong ≈40-bp stem, but with a 13-bp open loop (Fig. 4). A nick in this exposed loop piece can establish an accessible template for primer development and subsequent DNA replication. EBNA 1 binding helps to expose this special loop structure by binding stably to the stem part, leaving the single-stranded loop available to nicking. It is for this reason that we prefer the longer hairpin of Fig. 4B to the shorter one used by others (Fig. 4A), which in any case has an unstable short loop of only 3 bp. It would seem desirable to ascertain whether the presence of some but not all, say, two of the four, partial repeat copies in region II suffices for ori-P function. Alternatively, could deletion of the 13-bp loop destroy ori-P function? Also, experimental techniques are available by which to identify nicks or cleaved strands. In this way, it may be possible to ascertain whether nicks are readily produced in the loop piece of the region II repeats.

In view of the foregoing discussions, we might suggest that the region I repeats perform mainly as a site for strong EBNA 1 binding, whereas region II may be the key site of initiation of latent replication, as suggested earlier by Rawlins et al. (1985). EBNA 1 proteins are distinguished in having, bracketing a long 240-amino acid uncharged domain (exclusively glycine, alanine), peptide domains of 40–80 residues that alternate in being highly positively charged, then highly negatively charged (Table 5). This structure allows these proteins greater flexibility for simultaneous binding to region I and region II, and also in maintaining other protein–protein attractions. Furthermore, the dimeric nature of EBNA 1 (Luka et al. 1984) would permit the simultaneous binding to both regions I and II of the portion of it found by Rawlins et al. (1985) to bind to each of regions I and II. Alternatively, the existence of these two binding locations can be useful because binding proteins at one site may increase binding at the second site through cooperative interactions. Moreover, the nature of the physical binding can be fundamentally different at the region I repeats contrasted with the region II repeats, owing to their differences in copy numbers, length of unit, and spacings. In light of the results of Reisman and Sugden (1986), we may even suggest that region I may act as an activator for the action of palindromic region II, being, or being near, the origin of latent replication. Parenthetically, there are documented examples of binding proteins composed of at least two domains, one acting in a primary binding capacity and another for purposes of interactions with other proteins [e.g., λ repressor in λ-phage (Pabo and Lewis 1982), and the trp repressor in *Escherichia coli* (Schevitz et al. 1985)].

Acknowledgments. We acknowledge constructive comments from Michelle Calos, Alan Campbell, Ghassan Ghandour, Edward Mocarski, Robert Schimke, and an anonymous reviewer. This work was supported in part by NIH Grant GM10452-22, NIH Grant 5HL30856-02, and NSF Grant MCS82-15131 to S.K. and NIH Grant R01GM3056-01 to B.E.B.

References

Baer R, Banbier AT, Biggin MD, Deiniger PL, Farrell PJ, Gibson TJ, Hatfull G, Hudson GS, Satchwell SC, Sequin C, Tuffnell PS, Barrell BG (1984) DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature (London) 310:207–211

Cheah MSC, Ley TJ, Tronick SR, Robbins KC (1986) fgr proto oncogene mRNA induced in B-lymphocytes by Epstein-Barr virus infection. Nature (London) 319:238–240

Chow LT, Davidson N, Berg D (1974) Electron microscope study of the structures of λ dv DNAs. J Mol Biol 86:69–89

Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, DeRiel JK, Forget BG, Weissman SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Shoulders CC, Proudfoot NJ (1980) The structure and evolution of the human β-globin gene family. Cell 21:653–668

Fraenkel-Conrat H, Kimball PC (1982) Virology. Prentice-Hall, Englewood Cliffs NJ, p 182

Gray CP, Kaerner HC (1984) Sequence of the putative origin of replication of the U_L region of herpes simplex virus type 1 ANG DNA. J Gen Virol 65:2109–2119

Hauswirth WW, Berns AK (1979) Adeno-associated virus DNA replication: nonunit length molecules. Virology 93:57–68

Heller M, Flemington E, Kieff E, Deininger P (1985) Repeat arrays in cellular DNA related to the Epstein-Barr virus IR3 repeat. Mol Cell Biol 5:457–465

Henderson AS, Ripley M, Heller M, Kieff E (1983) Chromosome site for Epstein-Barr virus DNA in a Burkitt tumor cell line and in lymphocytes growth transformed *in vitro*. Proc Natl Acad Sci USA 80:1987–1991

Karlin S, Ghandour G (1985) Comparative statistics for DNA and protein sequences: single sequence analysis. Proc Natl Acad Sci USA 82:5800–5804

Klein HL, Petes TD (1981) Intrachromosomal gene conversion in yeast. Nature (London) 289:144–148

Luka J, Kreofsky T, Pearson GR, Hennessy K, Kieff E (1984) Identification and characterization of a cellular protein that cross reacts with the Epstein-Barr virus nuclear antigen. J Virol 52:833–838

Lupton S, Levine AJ (1985) Mapping genetic elements of Epstein-Barr virus that facilitate extrachromosomal persistence of Epstein-Barr virus-derived plasmids in human cells. Mol Cell Biol 5:2533–2542

Lusky M, Botchan MR (1984) Characterization of the bovine papilloma virus plasmid maintenance sequences. Cell 36:391–401

Miller G (1985) Epstein-Barr virus. In: Fields BN, Knipe DM, Melnick JL, Chanock RM, Roizman B, Shope RE (eds) Virology. Raven Press, New York, p 563–589

Milman G (1986) Sequence specific binding of Epstein-Barr virus nuclear antigen (EBNA 1). J Cell Biochem Suppl 10A: 216

Pabo CO, Lewis M (1982) The operator binding domain of λ repressor: structure and DNA recognition. Nature (London) 298:443–447

Petes TD (1980) Unequal meiotic recombination within tandem arrays of yeast ribosomal DNA genes. Cell 19:765–774

Rawlins DR, Milman G, Hayward SD, Hayward GS (1985) Sequence-specific DNA binding of the Epstein-Barr virus nuclear antigen (EBNA 1) to clustered sites in the plasmid maintenance region. Cell 42:859–868

Reisman D, Sugden B (1986) trans activation of Epstein-Barr viral transcriptional enhancer by the Epstein-Barr viral nuclear antigen 1. Mol Cell Biol 6:3838–3846

Reisman D, Yates J, Sugden B (1985) A putative origin of replication of plasmids derived from Epstein-Barr virus is composed of two *cis*-acting components. Mol Cell Biol 5: 1822–1832

Schevitz RW, Otwinowski Z, Joachimiak A, Lawson CL, Sigler PP (1985) The three-dimensional structure of trp repressor. Nature (London) 317:782–786

Shen S, Slightom JL, Smithies O (1981) A history of the fetal globin gene duplication. Cell 26:191–203

Streisinger G, Okada Y, Emrich J, Newton J, Tsugita A, Terzaghi E, Inouye M (1966) Frameshift mutations and the genetic code. Cold Spring Harbor Symp Quant Biol 31:77–84

Tautz D, Trick M, Dover GA (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature (London) 322: 652–656

van Santen V, Cheung A, Kieff E (1981) Epstein-Barr virus VII: size and direction of transcription of virus-specified cytoplasmic RNAs in a transformed cell line. Proc Natl Acad Sci USA 78:1930–1934

Waldeck W, Rösl F, Zentgraf H (1984) Origin of replication in episomal bovine papilloma virus type 1 DNA isolated from transformed cells. EMBO J 3:2173–2178

Watson JP, Tooze J, Kurtz DT (1983) Recombinant DNA, a short course. Scientific American Books, New York, p 131

Yates JL, Warren N, Reisman D, Sugden B (1984) A *cis*-acting element from the Epstein-Barr viral genome that permits stable replication of recombinant plasmids in latently infected cells. Proc Natl Acad Sci USA 81:3806–3810

Yates JL, Warren N, Sugden B (1985) Stable replication of plasmids derived from Epstein-Barr virus in various mammalian cells. Nature (London) 313:812–815

Received July 28, 1986/Revised December 31, 1986