Volume 17 Number 7 1989 Nucleic Acids Research

Overlapping redundant septuplets identical with regulatory elements of HIV-1 and SV40

Marian H.Seto*, Terence K.Brunck and R.L.Bernstein[1]

Triton Biosciences Inc., Department of Protein Chemistry, 1501 Harbor Bay Parkway, Alameda, CA 94501 and [1]San Francisco State University, Department of Biology, 1600 Holloway Avenue, San Francisco, CA 94132, USA

Received November 10, 1988; Revised and Accepted March 1, 1989

ABSTRACT

Overlapping redundant short oligomers in DNA sequences of retroviruses and papovaviruses have been identified. For each sequence, a search procedure determines the 5% short oligomers of the same length with the highest ratios of observed to expected occurrences based on singlet composition of the sequence. These short oligomers are referred to as compositionally-assessed redundant sequence elements (COARSEs). A pair of COARSEs overlapping by at least one base is considered to be a COARSE overlap. Most COARSE overlaps of the 7th order (overlapping septuplets) are found in long terminal repeats of retroviruses and in the regulatory control regions of papovaviruses SV40, BK and JC. Many of the 7th order COARSE overlaps in HIV-1 and SV40 are identical with regulatory elements determined experimentally. On the contrary, very few of the most frequently occurring oligomer overlaps, which are defined differently from COARSE overlaps, are present in the regulatory regions of retroviruses and papovaviruses. Examining DNA sequences of other genomes by the COARSE overlap method may identify putative regulatory regions.

INTRODUCTION

Regions of possible functional significance in DNA may be discovered by computer search algorithms which examine DNA sequences for linear patterns and other forms of ordering. A computer algorithm was written to locate overlapping redundant short oligomers in DNA sequences. The sequence length of the short oligomers considered is two to seven bases. For each sequence, the algorithm determines the 5% oligomers of the same length with the highest ratios of observed to expected occurrences based on singlet composition of the sequence. These oligomers are called compositionally-assessed redundant sequence elements (COARSEs). The algorithm then searches for all possible overlapping COARSEs without any mismatches, two at a time. A pair of COARSEs overlapping by at least one nucleotide is considered to be a COARSE overlap.

Retroviruses and papovaviruses SV40, BK, and JC are used here as the initial sequences to test the potential application of COARSE overlaps. Many regulatory elements controlling viral replication are located in long terminal repeats of retroviruses and in regulatory control regions of papovaviruses.

Since substantial biological information on the replication mechanisms of HIV-1 and SV40 is available, only HIV-1 and SV40 are analyzed in detail. In order to consider the information content of the entire viral genome, only complete genomic DNA sequences are used in this analysis.

An interesting question that arises from the COARSE overlap study is how COARSE overlaps differ from most redundant oligomer overlaps. For example, for each sequence, instead of selecting 5% oligomers of the same length that have the highest ratios of observed to expected occurrences, select the most redundant oligomers by sorting the frequencies of the oligomers. Then determine the 5% most frequently occurring oligomers, which we named most redundant sequence elements (MORSEs). Similar to COARSE overlaps, the number of overlapping MORSEs with no mismatches, two at a time, can be determined. A pair of MORSEs overlapping by at least one nucleotide is considered to be a MORSE overlap. Investigating retrovirus and papovavirus genomes by these methods shows that COARSE overlaps (but not MORSE overlaps) have potential for identifying regulatory regions.

METHODS

A computer algorithm, which we called OVERLAPsearch, determines COARSE overlaps in DNA sequences. Program OVERLAPsearch was written in Digital VAX-11 FORTRAN using GCG Procedure Library (1). The program finds the 5% identical-length oligomers with the highest ratios of observed to expected occurrences based on singlet composition. If selection of the exact 5% oligomers for a sequence is not possible, the percent nearest to but not greater than 5% is selected. The sequence length (or order) of short oligomers considered is two to seven bases. For each order, the program then identifies these short oligomers which overlap.

An example of results generated by program OVERLAPsearch is shown in FIGURE 1, which shows the 7th order COARSE overlaps in the -73 to -54 region of the 5' LTR of HIV-1. FIGURE 1A shows the septuplets within the region that are consider COARSEs - CGTGGCC, TGGCCTG, GCCTGGG, and CCTGGGC. These septuplets are among the 5% septuplets of the HIV-1 genome with the highest ratios of observed to expected occurrences. A COARSE overlap is a sequence of two overlapping COARSEs (FIGURE 1B). Since COARSEs of an overlap pair must have the same length and overlap by at least one base, the sequence length of a COARSE overlap is 8 to 13 bp long at the 7th order. Since there are four COARSEs in -73 to -54 region, there are six possible COARSE overlaps present in the region. FIGURE 1C shows a histogram indicating the location of COARSE

A) 7th order COARSEs (septuplets)

```
-73 CGTGGCCTGGGCGGGACTGG -54

    COARSEs
-73 CGTGGCC
-71   TGGCCTG
-69     GCCTGGG
-68      CCTGGGC
```

B) COARSE overlaps

Beginning[1] Position (5')	COARSEs	COARSE overlaps[2]	Total
-73	CGTGGCC TGGCCTG GCCTGGG CCTGGGC	CGTGGCCTG CGTGGCCTGGG CGTGGCCTGGGC	3
-71	TGGCCTG GCCTGGG CCTGGGC	TGGCCTGGG TGGCCTGGGC	2
-69	GCCTGGG CCTGGGC	GCCTGGGC	1

[1] First base position of the 5' end of the COARSE overlap.
[2] Underlined bases are bases common to both COARSEs.

C) Example of histogram showing COARSE overlaps

```
    3 2 1
-73 CGTGGCCTGGGCGGGACTGG -54
```

FIGURE 1. A) COARSEs in the -73 to -54 region of the 5' LTR of HIV-1 (HIVHXB2CG) - CGTGGCC, TGGCCTG, GCCTGGG, and CCTGGGC. These septuplets are among the 5% septuplets of the HIV-1 genome with the highest ratios of observed to expected occurrences based on singlet composition. B) A COARSE overlap is a sequence of two overlapping COARSEs. Since COARSEs of an overlap pair must have the same length and overlap by at least one base, the sequence length of a COARSE overlap is 8 to 13 bp long at the 7th order. There are 6 possible COARSE overlaps for the sequence shown in part A. Each overlap is a distinct pair of overlapping COARSEs. C) The number above a nucleotide indicates the number of occurrences of COARSE overlaps, with the beginning nucleotide of the 5' end of each COARSE overlap starting at that position. Only COARSEs which overlap are shown in these histograms. Non-overlapping COARSEs are not depicted in the histograms of FIGURES 2-8.

overlaps in the sequence. The number above a nucleotide indicates the number of occurrences of COARSE overlaps, with the first nucleotide of the 5' end of each COARSE overlap starting at that position. There are three COARSE overlaps at position -73. Each overlap is a distinct pair of septuplets overlapping by at least one nucleotide (FIGURE 1B). Two bases downstream, there are two COARSE overlaps; and two bases further downstream, there is one overlap. All six overlaps involve the four COARSEs shown in FIGURE 1A. Only COARSE overlaps are shown in the histograms used for the COARSE overlap analysis. COARSEs

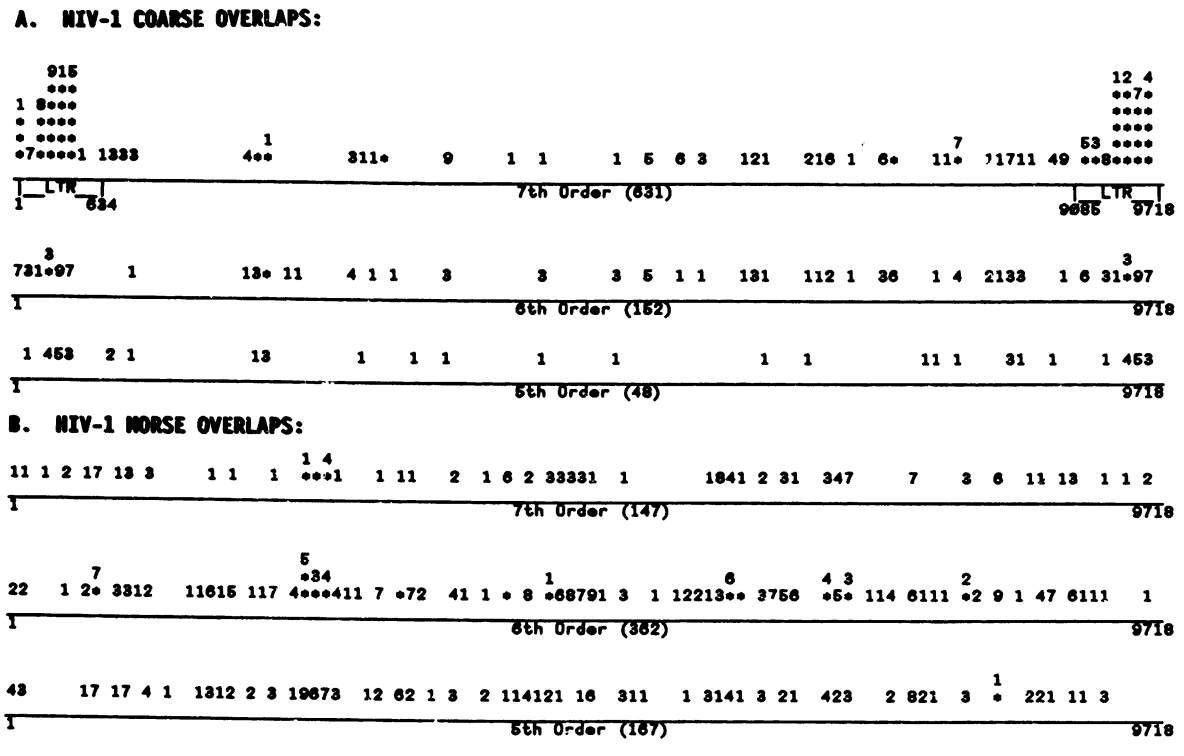

FIGURE 2. Each line represents the entire HIV-1 (HIVHXB2CG) genomic DNA sequence, with the first and last numbers below each line representing the beginning and ending positions of the genome. The number above the sequence represents the number of COARSE overlaps or MORSE overlaps at that position in the genome. The asterisk,*, indicates 10 overlaps. The number in parentheses indicates the total number of COARSE overlaps for the corresponding order.

which are not overlapping are not shown in the histograms.

A modified version of OVERLAPsearch was used to look for overlapping most redundant sequence elements (MORSEs). Similar to the method used to determine COARSEs, the total number of oligomers considered to be the most redundant sequence elements are among the 5% of the total possible identical-length oligomers of the DNA sequence. If selection of the exact 5% most redundant oligomers for a sequence is not possible, the percent nearest to but not greater than 5% is selected. The only difference between a COARSE and a MORSE is that a COARSE is determined by its ratio of observed to expected occurrences based on first-order nucleotide composition and a MORSE is determined by its number of occurrences. A sorting of the number of occurrences for all oligomers is required before MORSEs for a sequence can be determined.

Complete genomic DNA sequences from the GenBank viral database (release 55) were used in this analysis (2). The following papovaviruses were studied: Simian virus 40 (SV4CG), human JC polyomavirus (PLJCG), and two human papovaviruses BK - variants Dunlop (PVBDUN) and MM (PVBMM). Sequences of human and nonhuman retroviruses studied are shown in TABLE 1. For each genome, 500 randomized sequences were generated, each by permuting the bases of the original sequence 100 times. Therefore, the only difference between the random and the real sequences is the ordering of the bases. The average total number of COARSE overlaps was determined for the random sequences. Because the base composition of a real and its random sequence is the same, the different number of COARSE overlaps between the real and random sequences is determined by the ordering of bases. Randomizing the real sequences should give results predictable from the first-order base frequencies.

RESULTS

COARSE overlaps found in complete genomic viral sequences are shown in FIGURES 2A, 3A, and 4. The long solid line represents the entire DNA sequence and is scaled to the real sequence. The brackets below the solid line show where the 5' and 3' LTRs of retroviruses and the regulatory regions of papovaviruses are located. The numbers below the brackets for the retroviruses indicate the starting and ending positions of each LTR. The number above the sequence represents the number of COARSE overlaps at that position in the genome. The asterisk, *, indicates 10 overlaps. The region of the histogram with very high peaks represents extensive occurrences of COARSE overlaps. The

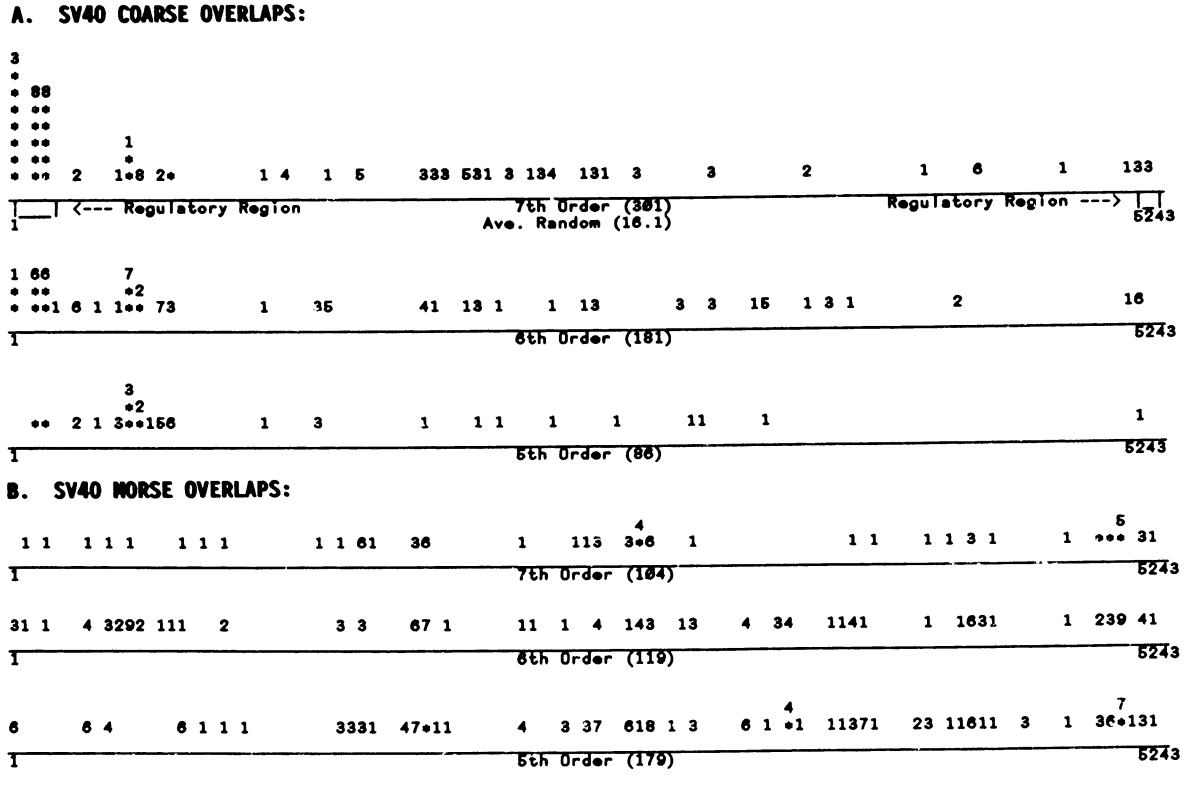

FIGURE 3. The locations of COARSE overlaps and MORSE overlaps are shown for SV40 (SV4CG). The histograms represent overlaps in identical manner as the histograms in Figure 2. Below the total 7th order COARSE overlap number for the real sequence is the average 7th order COARSE overlap number for 500 random sequences (shown in parentheses). The GenBank accession number for SV40 is J02400.

locations of MORSE overlaps for a genome are shown with a histogram described above (FIGURES 2B and 3B).

Retrovirus. For HIV-1, as the order increases, the total number of COARSE overlaps increases from 48 to 631 (FIGURE 2A). In each order, most COARSE overlaps occur prominently in 5' and 3' LTRs. Since there are 241 7th order COARSE overlaps in each LTR, the histogram has a distinctive appearance. The MORSE overlap histograms (FIGURE 2B) do not have the same appearance. Unlike COARSE overlaps, the number of MORSE overlaps does not increase as the order increases. There are 167 5th order, 362 6th order, and 147 7th order MORSE overlaps. Also unlike COARSE overlaps, most MORSE overlaps are not located in either the 5' or 3' LTRs. For all orders, the number of MORSE overlaps in a LTR ranges from 5 to 10.

Results for other human retroviruses, which include human T-cell lymphotropic virus I and II, are shown in TABLE 1A. The total number of 7th order COARSE overlaps for a human retrovirus exceeds 400, with at least 31% occurring in each LTR. Although the total number of 7th order COARSE overlaps for a nonhuman retrovirus varies more (TABLE 1B), from a minimum of 187 for Rous sarcoma virus to 676 for Simian Mason-PFizer D-type retrovirus, most COARSE overlaps remain in the LTRs. The percentage of total COARSE overlaps in LTRs ranges from 18% for Rous sarcoma virus to 50% for recombinant avian retrovirus MH2E21.

For random sequences of human retroviruses, the average total number of 7th order COARSE overlaps does not exceed 47.8, which is only approximately 10% of the total COARSE overlap number for a real sequence. The average total number of COARSE overlaps for a random nonhuman retrovirus ranges from 0.1 to 81.7, which is considerably less than the number for a real sequence.

Papovaviruses. The locations of 5th to the 7th order MORSE overlaps and COARSE overlaps for SV40 are shown in FIGURES 3A-B. Similar to HIV-1, as the order increases, the total number of COARSE overlaps increases from 86 to 301. Also similar to HIV-1, most COARSE overlaps occur extensively in the regulatory region of the genome at high orders, especially the 7th order. The results for MORSE overlaps differ from those of the COARSE overlaps. As the order increases from the 5th to the 7th, the total number of MORSE overlaps decreases from 179 to 104. Most MORSE overlaps are not located in the regulatory regions. BKV-Dunlop, BKV-MM, and JCV also have many 7th order COARSE overlaps in their respective regulatory regions (FIGURES 4A-C). Similar to SV40, MORSE overlaps are not present in the regulatory regions of BKV and JCV genomes (data not shown). Again, like retroviruses, the average total

TABLE 1. 7th Order COARSE overlaps in LTRs of Retroviruses

Retrovirus (complete genomic DNA sequence with entire 5' & 3' LTRs)	Database Entry Name - VIRAL (Accession Number)	Total (Real)	Number of COARSE overlaps Ave. Total (Random)[1]	5' LTR (Real)[2]	3' LTR (Real)[2]
A. Human Retroviruses					
Human Immunodeficiency virus-1	HIVHXB2CG (K03455)	631	36.0	241(.38)	240(.38)
Human Immunodeficiency virus-1	HIVPV22 (K02083)	508	18.4	184(.36)	184(.36)
Human Immunodeficiency virus-1	HIVSF2CG (K02007)	508	25.3	183(.36)	183(.36)
Human T-cell lymphotropic virus I	HL1PROP (J02029)	499	47.8	167(.33)	166(.33)
Human T-cell lymphotropic virus II	HL2CG (M10060)	420	28.1	130(.31)	130(.31)
B. Non-Human Retroviruses					
Recombinant avian retrovirus MH2E21	AC2E21CG (M14008)	328	1.0	163(.50)	163(.50)
Fujinami sarcoma virus	ACF (J02194)	441	30.6	208(.47)	207(.47)
Rous sarcoma virus	ALRCG (J02343)	187	21.2	34 (.18)	34 (.18)
Bovine leukemia virus	BLV (K02120)	286	18.7	78 (.27)	69 (.24)
Equine infectious anemia virus	EIAV (M16575)	414	81.7	159(.38)	159(.38)
Mouse mammary tumor virus	MMTPROCG (M15122)	365	9.5	96 (.26)	96 (.26)
FBJ murine osteosarcoma virus	MSJMUSV (J02084)	763	0.1	362(.47)	362(.47)
Moloney murine sarcoma virus	MSMPROCG (J02266)	510	6.1	226(.44)	217(.43)
FBR murine osteosarcoma virus	MSRMUSV (K02712)	251	0.5	98 (.39)	98 (.39)
Simian Mason-Pfizer D-type retrovirus	SIVMPCG (M12349)	676	10.3	194(.29) 189(.28)[3]	191(.28)
Simian SRV-1 type D retrovirus	SIVRV1CG (M11841)	497	36.8	174(.35)	176(.35)

[1] Random = Average total number of COARSE overlaps for 500 randomized sequences, each generated by permuting the bases of the original sequence 100 times.

[2] The left number is the number of COARSE overlaps in each LTR of the real sequence. The number in parentheses indicates the percentage of COARSE overlaps in each LTR [actual no. COARSE overlaps per LTR/total no. COARSE overlaps per genome].

[3] Simian Mason-Pfizer D-type retrovirus (Viral:sivmpcg) contains two 5' LTRs (347bp and 328bp).

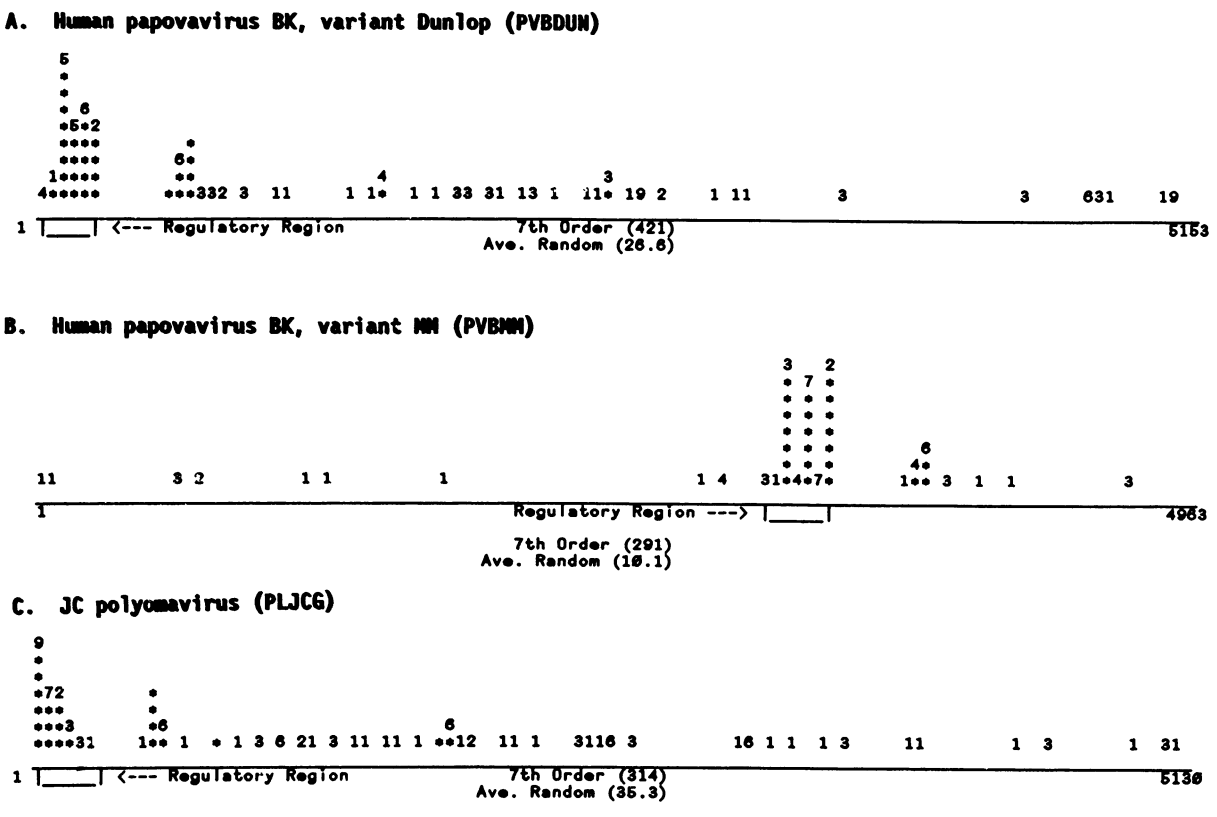

FIGURE 4. The locations of 7th order COARSE overlaps are shown for papovaviruses BK (Dunlop), BK (MM), and JC in histograms A, B, and C, respectively. The histograms represent overlaps as in Figure 2. The average COARSE overlap number is based on 500 random sequences. The GenBank accession numbers for BKV (Dunlop), BKV (MM), and JC are J02038, J02040, and J02226, respectively.

COARSE overlap number for random SV40, BKV, or JCV sequences is usually less than 10% of the number for its respective real sequence. The total 7th order COARSE overlap numbers for SV40, BK-Dunlop, BK-MM, and JC are 301, 421, 291, and 314, respectively. The average total 7th order COARSE overlap numbers for the respective random sequences are 16.1, 26.6, 10.1, and 35.3.

DISCUSSION

The occurrences of both MORSE overlaps and COARSE overlaps in HIV-1 and SV40 were separately examined. The results show that few MORSE overlaps occur in the long terminal repeats of HIV-1 and the regulatory region of SV40. On the contrary, most 7th order COARSE overlaps (overlapping septuplets) occur in these regions. Preliminary results indicate that 7th order COARSE overlaps, not MORSE overlaps, of these viruses correlate with important biological information. This analysis implies that short sequence elements which overlap may help to regulate integrational and transcriptional functions of these retroviruses and papovaviruses. Random sequences have few 7th order COARSE overlaps compared with their respective real sequences. Many of the COARSE overlaps of HIV-1 and SV40 have been experimentally shown to be identical with regulatory elements as described below.

HIV-1. Deletion analysis of the HIV LTR have been performed to determine regions of the LTR required for transcriptional activity in HeLa whole cell extracts (3). Results from the deletion analysis reveal that the region from -104 to -57 is required for HIV-1 LTR-directed transcription in vitro (FIGURE 5). The SV40 enhancer core consensus sequence and the Sp1 binding sites (4) within this region have been suggested to be important for activating transcription (3). The -104 to -57 region contains many COARSE overlaps. Also present in this region are two nuclear factor binding sequences, GGGACTTTCC, called NF-κB sites (5). COARSE overlaps occur extensively in these NF-κB binding sites. Studies indicate that this region responds to T-cell activation signals (5,6).

The trans-acting responsive region (TAR) between +1 and +80, which contains many COARSE overlaps, is capable of forming several secondary structures (3,7,8,9). One stable secondary structure type is shown in FIGURE 5 (3). The potential RNA hairpin structures have been suggested to affect the TAR function (7, 8, 9). There are further experimental results which indicate that COARSE overlap regions of HIV-1 LTR are indeed binding sites (10). HeLa cellular proteins, called EBP-1 and UBP-1, have been isolated. These two proteins bind to the GGGACTTTCC repeats of the enhancer element and CTCTCTGG repeats of the TAR region, respectively. The deletion of nucleotides between

-340 and -185, the negative regulatory element (NRE), increases the level of gene expression (11). This region has some COARSE overlaps at the start of the sequence, but has few COARSE overlaps overall. For the 5' LTR of HIV-1 genome, COARSE overlaps do occur in regions of the LTR which are required for transcriptional activity.

The search procedures used to locate MORSE overlaps and COARSE overlaps both identify repeating elements in DNA sequences. This trend is more apparent when the 3' LTR is removed from retroviruses (data not shown). For example, the total 7th order COARSE overlap number for the HIV-1 variant (HXB2) with the 3' LTR deletion decreases from 631 to 267, with 36 (13%) located in the 5' LTR. The total 7th order MORSE overlap number for the same HIV-1 variant without the 3' LTR decreases from 147 to 124, and only one is located in the 5' LTR. Even with the deletion of the 3' LTR, which is a repeating sequence of 634 bp in length, the results indicate that 7th order COARSE overlaps, not MORSE overlaps, are present in the regulatory regions of these viruses.

SV40. In SV40, both the initiation of viral DNA replication and the control of early and late gene transcription are mediated by the large T antigen. The large T Ag binds specifically to three sites within the regulatory region - I, II, III. Each contains repeats of sequence G/T-A/G-GGC (12). Experimental results suggest that the pentameric repeats of sequence G/T-A/G-GGC form the core of the recognition-binding sites for the large T Ag. All three binding sites contain COARSE overlaps (FIGURE 6).

The early and late RNA syntheses proceed in opposite directions from the regulatory region. The 21-bp repeat region, which contains six GC-motifs, regulates both the early and late gene expression (13,14). Each GC-motif is an independent binding site for the cellular transcription factor Sp1. Analyses of mutants with point mutations within the GC-motifs have shown that motifs I and II are important for the transcription of early genes (T antigens); and IV, V and VI are important for late genes (capsid proteins). COARSE overlaps are located in all six GC-motifs.

The 72 bp repeat region functions as an enhancer of transcription, composed of three discrete elements A, B, and C (15). COARSE overlaps occur extensively in the 72 bp repeats, with most of the COARSE overlaps situated between elements B and C. Element C, which contains the GTGG-A/T-A/T-A/T-G ('core' consensus), has some COARSE overlaps. Between element B and C are the TC motifs - TCCCCAG (16). The TC motifs are binding sites for a 52 kD nuclear enhancer binding protein, called AP-2, which activates early transcription of SV40 (17).

BKV. Within the GenBank database there are two complete BKV genomes: BK-MM

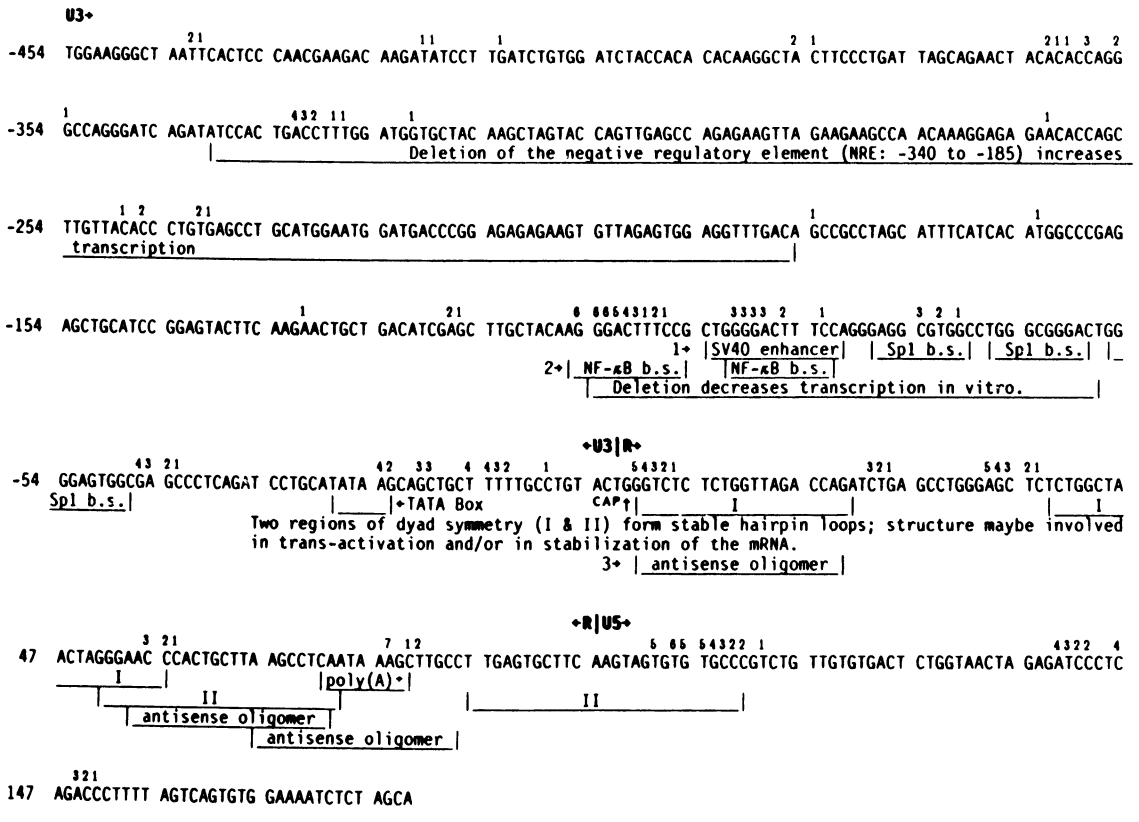

FIGURE 5. 7th order COARSE overlaps in 5' LTR (634 bp) of HIV-1 (HIVHXB2CG). 1-› Homology with the complementary strand of SV40 enhancer. 2-➧ b.s. =binding site. 3-➧ Antisense oligodeoxynucleotides inhibit HIV-1 replication.

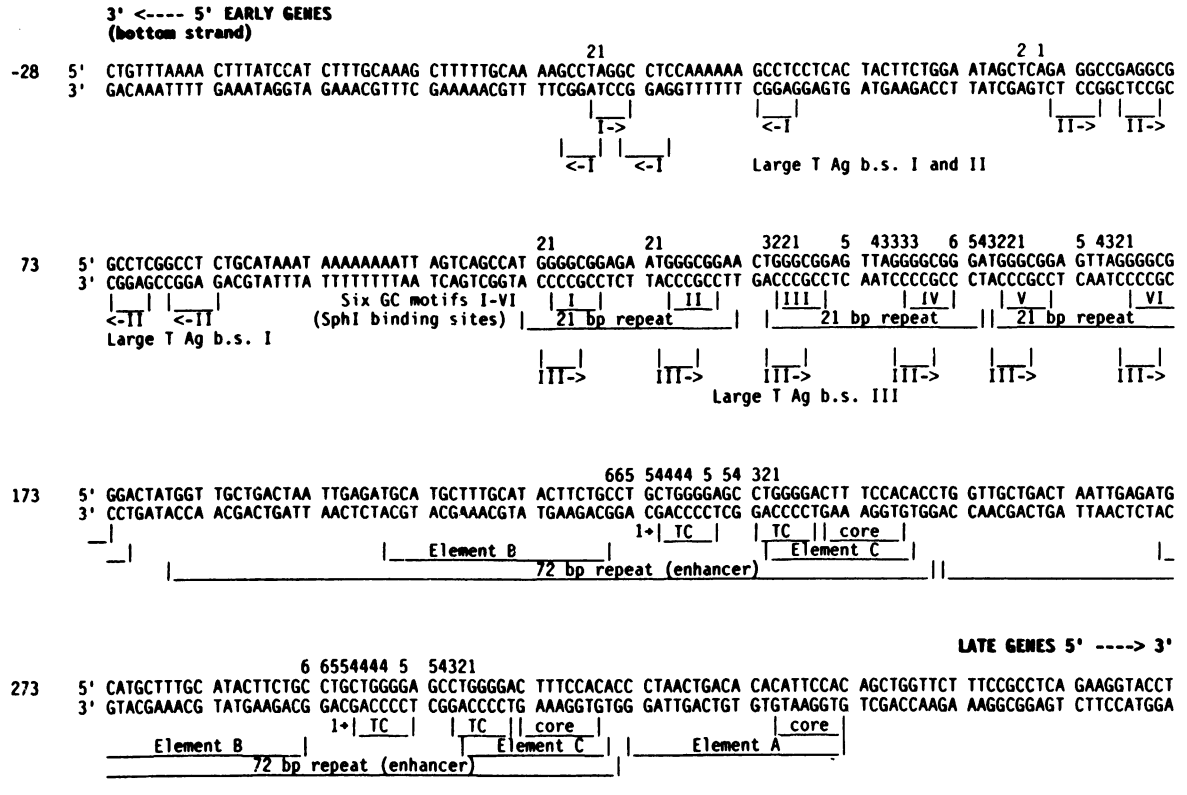

FIGURE 6. The locations of 7th order COARSE overlaps are shown for SV40 (SV4CG). 1→ TC = TC-motifs bind to 52 kD nuclear enhancer binding protein called AP-2. (b.s. = binding sites)

and BK-Dunlop. COARSE overlaps occur extensively in the regulatory regions of these two BKVs (FIGURE 4). For BK-Dunlop, all three large T Ag binding sites contain COARSE overlaps (FIGURE 7A) (18). Recognition sites for cellular proteins have been identified for some BK variants by the technique of DNAase I footprinting (19). BKV-Dunlop contains five sites, called N1, N2, N3, A1, and A2. Binding sites N1, N2, and N3 recognize nuclear factor NF-BK, a member of the nuclear factor I (NFI) family of transcription factors. Binding sites A1 and A2 recognize nuclear factor AP-1, a transcription factor for SV40. BK-MM contains six recognition sites for NF-BK (N1, N4, N7, N8, N9, N10), and a single binding site (S1) for transcription factor Sp1 (FIGURE 7B). COARSE overlaps occur extensively in binding sites N4, N8, and N10.

JCV. Sequence analyses of JCV variants indicate that the regulatory regions of these isolates are hypervariable due to complex alterations of the original 98 bp repeat of the wild type JCV, Mad-1 (20,21). Apart from the variations, there are conserved sequences in the regulatory regions of these isolates (FIGURE 1C). A 17 bp conserved sequence (region II) is present in naturally occurring infectious JCV variants (20). A longer conserved region (region I) common to JCV variants isolated from the central nervous system and the kidney of a patient contains putative enhancer elements (21). This conserved region, which includes the 17 bp sequence, contains many COARSE overlaps. Conserved sequences and novel enhancer core sequences, generated from DNA rearrangement, can influence the host ranges of JCV variants.

Regions with COARSE overlaps. Examining linear nucleic acid sequences of viral genomes by the COARSE overlap method may identify prospectively possible regulatory regions. For example, the regulatory regions of SV40, JCV, and BKV, which contain many COARSE overlaps, are vital to the propagation and replication mechanisms of these papovaviruses. Sequence elements within regulatory regions of SV40 and JCV contribute to host specificity and tissue tropism (22,23). DNA rearrangements within regulatory regions of JCV and BKV affect the host ranges of these viruses (21,24,25). This sequence analysis shows that COARSE overlaps occur extensively in LTRs of retroviruses. LTRs of murine and human retroviruses have been suggested to encode pathogenic determinants (26). From studies that were based mainly on recombinant viruses of leukemogenic SL3-3 (causes T-cell lymphomas in mouse cells) and nonleukemogenic Akv viruses (26,27), the sequences of the SL3-3 LTRs have been shown to encode some leukemogenic determinants. One of the leukemogenic recombinant viruses is the coding region of the original Akv with the SL3-3 LTRs (FIGURE 8). There are more COARSE overlaps occurring within the LTRs of the leukemogenic recombinant virus, Akv with SL3-3 LTRs, than the original

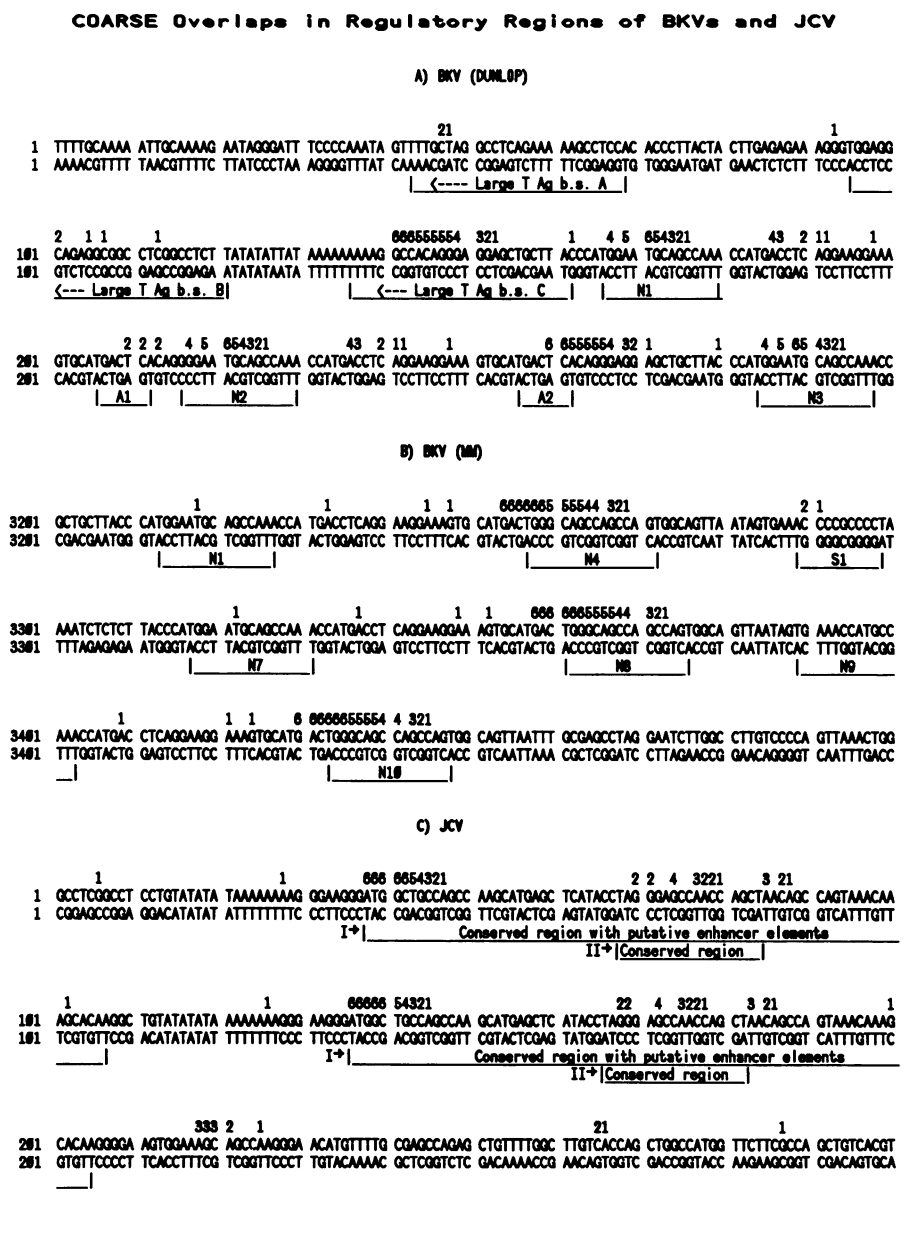

FIGURE 7. Late genes 5'--› 3' (top strand). Early genes: 3'‹--5' (bottom strand). A) b.s. = binding sites for large T Ag. A & B) N1, N2, N3, N4, N7, N8, N9, N10, A1, A2, and S1 = binding sites for nuclear factors. C) I and II = Conserved sequences with putative enhancer elements. Nucleotide position is based on GenBank database files.

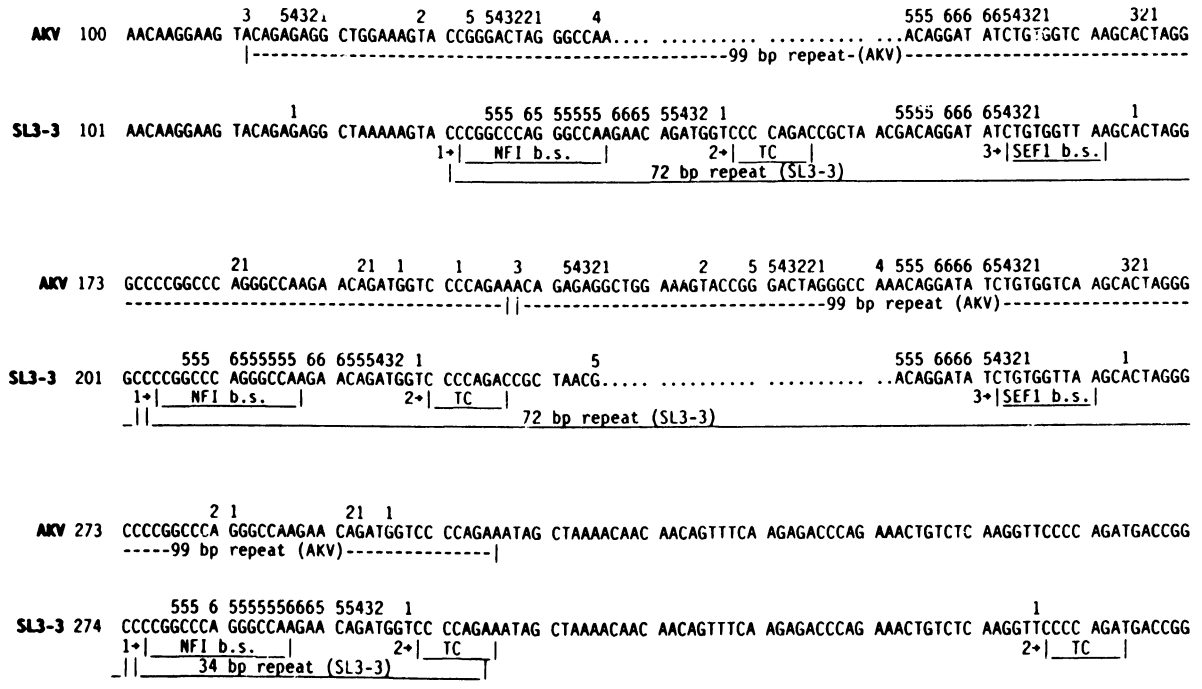

FIGURE 8. The 7th order COARSE overlaps are shown for Akv and a recombinant - Akv with SL3-3 LTRs. The COARSE overlap determination here is based on the entire Akv genome with complete 5' and 3' LTRs, under the assumption of near-identity between the left and right LTRs (the Genbank sequence itself is not complete). The SL3-3 LTRs differ form the Akv LTRs with respect to the arrangement of the tandem repeats. 1➔ NFI b.s. = binding site for nuclear factor I, which enhances adenovirus replication. 2➔ TC = homologous to TC-motifs of SV40, which are binding sites for nuclear factor AP-2. 3➔ SEF1 b.s. = binding site for SL3-3 enhancer factor 1 proteins.

Akv. Extensive occurrences of COARSE overlaps in the SL3-3 LTRs of the recombinant virus are located within binding sites for nuclear proteins NFI and SEFI (28,29). Experimental data suggest that some of these binding site motifs are the primary determinants of the T-cell tropism and leukemogenicity of the SL3-3 virus (28,29). COARSE overlaps present in the recombinant virus, especially those located in the LTRs of SL3-3 but not Akv, may contribute in part to the leukemogenicity of the SL3-3.

Potential of COARSE overlaps. If COARSE overlaps can be consistently shown to occur in important regions of other complete viral genomes, not only regulatory elements but also regions which encode essential proteins (trans-activating factor, adsorption protein), then the procedure for finding COARSE overlaps may be used as a tool for analysis of sequence information. COARSE overlaps may address the problem of selecting the most important targets of the viral genome to study. For example, synthetic anti-sense oligodeoxynucleotides have been shown to inhibit HIV-1 replication and gene expression in cultured human cells (30,31). Some of these synthetic oligomers are complementary to regions of the LTRs with COARSE overlaps. When twenty anti-sense oligodeoxynucleotides to HIV-1, which included 17 anti-sense oligomers not located in the LTRs, were ranked for activity based on inhibitions of syncytia formation (31), the most effective oligomers were from the R region of the LTRs - a 5' untranslated region (54 to 73) and poly(A)+ signal (9183 to 9202) (FIGURE 8). Both oligomers are 20 nucleotides in length and contain COARSE overlaps. Another very good anti-sense oligomer was targeted against the CAP site, which is identical with many COARSE overlaps (30).

In summary, many of the 7th order COARSE overlaps in the regulatory regions of HIV-1 and SV40 are identical with experimentally determined regulatory elements. On the contrary, very few of the MORSE overlaps are present in the regulatory regions of these viruses. The COARSE overlap analysis implies that overlapping compositionally-assessed redundant sequence elements may regulate replication and transcription for these viruses.

ACKNOWLEDGEMENTS

The authors would like to thank John Monahan for reviewing the manuscript.

•To whom correspondence should be addressed

REFERENCES

1. Devereux, J., Haeberli, P., Smithies, O. (1984) Nucleic Acids Res. 12, 387-395.
2. Bilosky, H.S., and Burks, C. (1988) Nucleic Acids. Res. 16, 1861-1863.

GenBank Sequence Data Bank - release 55, March 1988.
3. Patarca, R., Heath, C., Goldenberg, G.J., Rosen, C.A., Sodroski, J.G., Haseltine, W.A., and Hansen, U.M. (1987) Aids Res. and Human Retroviruses. 3, 41-55.
4. Jones, K.A., Kadonaga, J.T., Luciw, P.A., and Tjian, R. (1986) Science 232, 755-759.
5. Nabel, G. and Baltimore, D. (1987) Nature 326, 711-713.
6. Tong-Starksen, S.E., Luciw, P.A., and Peterlin, B.M. (1987) Proc. Natl. Acad. Sci. USA 84, 6845-6849.
7. Feng, S., and Holland, E.C. (1988) Nature 334, 165-167.
8. Muesing, M.A., Smith, D.H. and Capon, D.J. (1987) Cell 48, 691-701.
9. Okamota, T., Wong-Staal, F. (1986) Cell 47, 29-35.
10. Wu, F.K., Garcia, J.A., Harrich, D., and Gaynor, R.B. (1988) EMBO J. 7, 2117-2129.
11. Rosen, C.A., Sodroski, J.G., Haseltine, W.A. (1985) Cell, 41, 813-823.
12. Pomerantz, B.J., and Hassell, J.A. (1984) J. Virol. 49, 925-937.
13. Dynan, W.S., and Ayer, D. (1987) In Reznikoff, W.S., Burgess, R.R., Dahlberg, J.E., Gross, C.A., Record, M.T.R., and Wickens, M.P., (eds.) RNA polymerase and the regulation of transcription: proceedings of the sixteenth Steenbock symposium, Elsevier Science Publishing Co., New York, pp. 303-311.
14. Ernoult-Lange, M., Omilli, F., and May, E. (1987) Nucleic Acids Res. 15, 8177-8193.
15. Ondek, B., Shepard, A., and Herr, W. (1987) EMBO J. 6, 1017-1025.
16. Zenke, M., Grundström, T., Matthes, H., Wintzerith, M., Schatz, C., Wildeman, A. and Chambon, P. (1986) EMBO J. 5, 387-397.
17. Mitchell, P.J., Wang, C. and Tjian, R. (1987) Cell 50, 847-861.
18. Ryder, K., Delucia, A. L. and Tegtmeyer, P. (1983) Virology 129, 239-245.
19. Markowitz, R.-B. , and Dynan, W.S. (1988) J. Virol. 62, 3388-3398.
20. Martin, J.D, King, D.M., Slauch, J.M., and Frisque, R.J. (1985) J. Virol. 53, 306-311.
21. Loeber, G. and Dörries, K. (1988) J. Virol. 62, 1730-1735.
22. Laimins, L.A., Khoury, G., Gorman, C., Howard, B., and Gruss, P. (1982) Proc. Natl. Acad. Sci. USA 79, 6453-6457.
23. Kenney, S., Natarajan, V. Strike, D., Khoury, G., and Salzman, N.P. (1984) Science 226, 1337-1339.
24. Miyamura, T., Furuno, A., and Yoshiike, K. (1985) J. Virol. 54, 750-756.
25. Watanabe, S., Soeda, E., Uchida, S., Yoshiike, K. (1984) J. Virol. 51, 1-6.
26. Sodroski, J., Patarca, R., Lenz, J., Trus, M., Crowther, R., Perkins, D., Gallo, R.C., Wong-Staal, F., Josephs, S., Gelmann, E.P., Haseltine, W.A. (1984) In Gallo, R.C., Essex, M.E., and Gross, L. (eds.) Human T-cell leukemia/lymphoma virus. Cold Spring Harbor Laboratory Press, New YorK, pp. 149-155.
27. Lenz, J., Celander, D., Crowther, R.L., Patarca, R., Perkins, D.W., and Haseltine, W. A. (1984) Nature 308, 467-470.
28. Hallberg, B. and Grundström, T. (1988) Nucleic Acids. Res. 16, 5927-5944.
29. Thornell, A., Hallberg, B., and Grundström, T. (1988) Mol. Cell. Biol. 8, 1625-1637.
30. Zamecnik, P.C., Goodchild, J., Taguchi, Y., Sarin, P.S. (1986) Proc. Natl. Acad. Sci. USA 83, 4143-4146.
31. Goodchild, J., Agrawal, S., Civeira, M.P., Sarin, P.S., Sun, D., and Zamecnik, P.C. (1988) Proc. Natl. Acad. Sci. USA 85, 5507-5511.