JOURNAL OF VIROLOGY, July 1983, p. 137–145
0022-538X/83/070137-09$02.00/0

Vol. 47, No. 1

Provirus of M7 Baboon Endogenous Virus: Nucleotide Sequence of the *gag-pol* Region

TAKA-AKI TAMURA

Department of Microbiology, Keio University School of Medicine, Shinjuku-ku, Tokyo, Japan-160

Received 1 November 1982/Accepted 31 March 1983

A 3,023-base nucleotide sequence of the M7 baboon endogenous virus genome, spanning the 5′ noncoding region as well as the entire *gag* gene and part of the *pol* gene, is reported. Within the 562-base 5′ noncoding region, a 21-base sequence complementary to the OH terminus of $tRNA^{pro}$ is located immediately downstream from the long terminal repeat. Amino acid sequences were deduced from the 1,596 nucleotides comprising the *gag* gene, and the four structural *gag* polypeptides, p12, p15, p30, and p10, appeared to be coded contiguously. Only one termination codon interrupted the M7 *gag* and *pol* genes. The data suggest that 55 additional amino acids may be attached to the NH_2 terminus of the *gag* precursor protein. However, such a sequence was not detected in virions or in virus-infected cells. With the exception of the p15 region, nucleotide and amino acid sequences of the *gag* and *pol* regions of M7 virus exhibited strong homologies to those of Moloney leukemia virus.

In replication-competent mammalian retroviruses, gene organization of the integrated proviral DNA (5′-LTR-*gag-pol-env*-LTR-3′), where LTR is the long terminal repeat, is colinear with viral genomic RNA (3, 17). The 5′ end of the integrated proviral DNA encodes several functions necessary for replication of the viral genome and synthesis of viral proteins (3, 31, 35, 46). The *gag* gene of mammalian retroviruses codes for four core proteins, whose functions are not yet completely characterized (12, 22, 35, 38).

Noda et al. and Tamura et al. have previously described cloning of the 8.2-kilobase-pair (kb) proviral DNA of M7 baboon endogenous virus and have reported the sequence of the 554-nucleotide-long LTR of the provirus (23, 41). Signals required for initiation and termination of transcription of the viral genome were found to be present in the LTR.

In this report I describe the nucleotide sequence from the LTR through the *gag* gene and into the proximal region of the *pol* gene. The 5′ noncoding region of the provirus contains sequences for initiation of reverse transcription, splicing of viral RNA, and initiation of protein synthesis. In the *gag* and *pol* regions of M7, nucleotide and amino acid sequences exhibit considerable homology with corresponding regions of Moloney murine leukemia virus (Mo-MuLV) proviral DNA (37). Functionally important regions appear to have been conserved throughout long evolutionary periods. Interestingly, however, the sequence of the p15 region of the M7 *gag* gene does not seem to bear a close relationship to the analogous region (p12) of Mo-MuLV.

MATERIALS AND METHODS

Recombinant DNA. Circular M7 DNAs were cloned into the phage vector Charon 28 (23). All analyses were performed on fragments of the M7 DNA which were derived from a recombinant phage, λBEV-11, and subsequently cloned in pBR322. Detailed procedures for preparation of the DNA fragments and construction of the recombinant plasmids have been described previously (41). Maps of the M7 DNA in λBEV-11 and the subcloned fragments are shown in Fig. 1.

Sequence analysis. I used three subcloned M7 DNAs, pBE-B8, pBE-SH, and pBE-L, as sources for sequencing (Fig. 1) (41). Fine-structure maps of this region and the sequencing strategies employed are shown in Fig. 2 A and B. The chemical modification methods for DNA sequencing used in these studies were essentially those of Maxam and Gilbert (21). Detailed conditions for chemical modification and electrophoresis have been described previously (41). After electrophoresis, the gels were dried, overlaid on Kodak XAR-5 or XRP-1 X-ray film, and then exposed at −40 or −80°C for appropriate periods.

RESULTS

Sequencing strategy. Sequences downstream from the 3′ end of the M7 baboon virus LTR, extending through the entire *gag* gene and into the *pol* region, were analyzed. These sequences were derived from three different recombinant plasmids. The first, pBE-B8, contained two non-

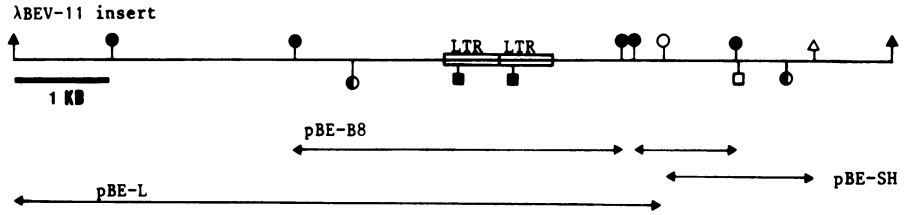

FIG. 1. Restriction map of provirus DNA cloned in λBEV-11 and fragments contained in subclones used in this study. Sites for *Eco*RI (▲), *Bam*HI (●), *Bgl*II (◐), *Xho*I (■), *Sal*I (○), *Pst*I (□), and *Hin*dIII (△) are indicated.

contiguous *Bam*HI fragments, a 3.2-kb fragment containing tandem LTRs and a second 1.0-kb fragment (Fig. 1) (41). These two fragments were situated in the proviral genome such that they bracketed a 160-base-pair (bp) *Bam*HI fragment not present in pBE-B8. The sequence of this 160-bp fragment was derived from pBE-L, a second recombinant plasmid which carries an *Eco*RI-*Sal*I fragment encompassing the smaller sequence (Fig. 1). Based on the known sizes of other retroviral *gag* genes (13, 37, 40, 45), that of M7 should extend from the LTR to a point no further in the 3′ direction than the single *Hin*dIII site shown in Fig. 1. Thus, analysis of a third recombinant plasmid, pBE-SH, which contains a 1.5-kb fragment from the *Sal*I to the *Hin*dIII site, should be sufficient to complete the sequence of the M7 *gag* gene. Detailed strategies for sequencing are summarized in Fig. 2B. In Fig. 2A, the sites of four restriction endonucleases as determined by sequence analysis are also shown.

Nucleotide sequence of the *gag-pol* region of M7 provirus. The nucleotide sequence of 2,469 bases, extending from the 3′ end of the M7 to the region near the unique *Hin*dIII site, is shown in Fig. 3. The nucleotides are numbered from the first T residue just downstream from the LTR and represent the positive strand, which corresponds to the viral genomic RNA. The LTR sequence is also presented and numbered negatively from 3′ end. Functionally important regions for transcription of the viral RNA (CAT box, TATA box, polyadenylation and termination signals, and a putative capping site) are shown in Fig. 3 and have been described previously (41). The sites of several restriction endonucleases which recognize 4- or 6-bp sequences are shown in Fig. 2 and 3. A potential target sequence for recombination within the *gag* region, ATAA (position 1,780), has been previously described (42). Finally, a 21-bp stretch of DNA, from positions 1 to 21, is complementary to the OH terminus of $tRNA^{pro}$.

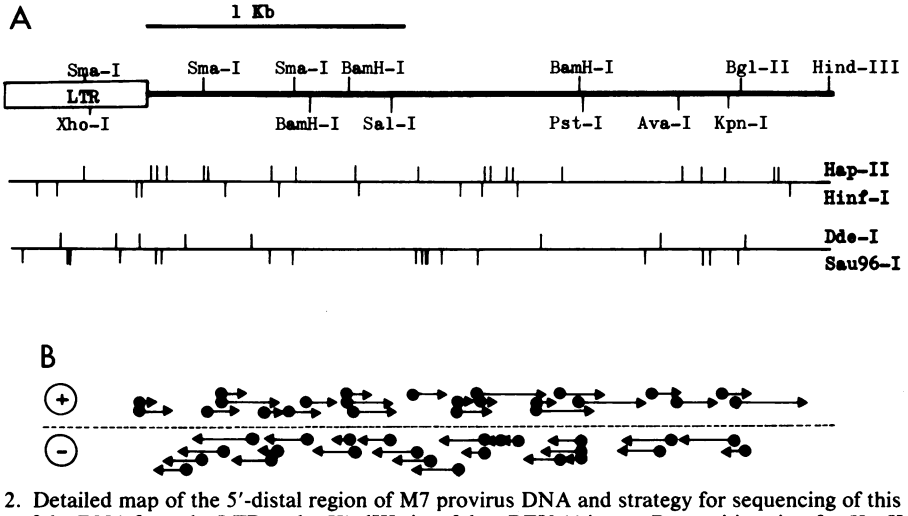

FIG. 2. Detailed map of the 5′-distal region of M7 provirus DNA and strategy for sequencing of this region. (A) Maps of the DNA from the LTR to the *Hin*dIII site of the λBEV-11 insert. Recognition sites for *Hap*II, *Hin*fI, *Dde*I, and *Sau*96I as determined by sequencing are indicated by vertical lines. (B) Arrows, whose positions correspond to the maps shown in (A), indicate DNA stretches determined in this study. +, Same polarity as that of the viral genome.

Open reading frames for M7 *gag* protein. To determine the amino acid structure of the M7 *gag* gene product, I searched for open sequences in all three reading frames in the nucleotides downstream from the putative capping site at position −133 (Fig. 3; Table 1). There are numerous ATG initiation codones. The three termination codons, amber (TAG), opal (TGA), and ochre (TAA), appear at many sites in reading frames 2 and 3, but only rarely in reading frame 1. Therefore, only reading frame 1 is able to encode a significantly large polypeptide. Accordingly, only the amino acid sequence deduced from reading frame 1 is shown in Fig. 3.

Several lines of evidence have suggested that the order of polypeptides in the M7 *gag* polyprotein is NH_2-p12-p15-p30-p10-COOH (2, 25, 39). In addition, Copeland et al. (8) have determined the amino and carboxy termini of thes M7 *gag* peptides. From these previous data, I was able to determine the termini of each *gag* peptide and have indicated these in Fig. 3. The amino acid sequences predicted from the nucleotide sequence shown in Fig. 3 are generally consistent with the findings of Copeland et al., although mismatchings of several amino acids were observed.

If the ATG codon at position 262 is assumed to be the initiation site for *gag* gene translation, then the calculated molecular weight of the peptide coded by this sequence should be approximately 12,000. In addition, the base sequences in the NH_2 terminus of the M7 *gag* seem to be partially homologous to those of Mo-MuLV (see Fig. 4). Hence, I assumed that the ATG codon at position 430 is the translation initiation codon for the M7 *gag* protein and that the open reading frame comprising the entire M7 *gag* protein is 1,596 bases in length. The region from the postulated capping site at positions −133 to +429 is, therefore, considered to be the 5′ noncoding region of this virus. The identities of the first and last amino acid residues of the *gag* polyprotein could not be determined with certainty in this study, since such *gag* precursor polyproteins may lose several amino acids from both amino and carboxy termini during processing (12, 18, 37, 47). The structure of the mature M7 *gag* proteins are discussed in more detail below.

A second long open reading frame was found in the DNA stretch immediately downstream from the amber termination codon of the gag gene at position 2,026. This second open reading frame would appear to code for the *pol* protein.

DISCUSSION

Primer binding site for reverse transcription. Retroviral RNA is transcribed into DNA by reverse transcriptase, whose primer is a cellular tRNA which binds to a site near the 5′ end of the viral genome (27, 32, 43). In an earlier publication (41), I reported that the 14 bases at the end of the M7 LTR could hybridize with the OH terminal sequences of $tRNA^{pro}$. In the present study, I confirmed the previous findings and showed that the $tRNA^{pro}$ binding site actually consists of 21 bases, positions 1 to 21 (Fig. 3). In murine retroviruses (29, 37, 44, 45), 17 bases have been identified as the primer binding site for $tRNA^{pro}$.

Sequences for splicing in the 5′ noncoding region. Subgenomic mRNAs of retroviruses have been detected which represent spliced products consisting of a leader sequence derived from the 5′ end of the viral genome joined to the coding regions of the *env* or *src* gene (16, 30, 48). Subgenomic 20S RNA has also been detected in M7-infected human cells (unpublished data). The 5′ noncoding region of the viral genome was observed to harbor signal sequences for splicing. One DNA stretch, TCAGGTACT, was located at positions 122 to 130 in the middle of the 5′ noncoding region and resembled the consensus sequences for the splicing donor, AG/GTRAG or TCAG/C (5, 36). M7 RNA may be spliced from this site to the 5′ end of the *env* coding region. No sequences for splicing were detected in the LTR. It has been suggested that Mo-MuLV and Moloney murine sarcoma virus each have a splicing donor sequence about 70 bp downstream from the LTR (29, 37, 45). In the case of Rous sarcoma virus (9, 15, 40), the region from the capping site to the donor site preceding the *gag* gene has been shown to be the leader sequence of the viral RNA. Acceptor sequences for splicing are located at positions 56 and 260, although the significance of the presence of these sequences in the 5′ noncoding region is not clear at present.

Initiation site for translation of the *gag* gene. As described above, three ATG codons were detected at positions 262, 402, and 410 in the region preceding the *gag* gene (Fig. 3), although the actual initiation codon for the *gag* precursor polyprotein of the M7 virus is located at position 430. These findings are unexpected; in eucaryotic cells, ribosomes bind at or near the 5′ end of mRNA and then "travel" downstream to the first AUG codon, where translation is initiated (20). Similar, apparently superfluous, initiation codons have, however, been detected in other viral RNAs including Mo-MuLV (19, 37, 40). A glycosylated *gag* precursor of Mo-MuLV, $gPr80^{gag}$, which is larger than the major precursor of the *gag* polyprotein, Pr65, has been detected previously (11, 33). This $gPr80^{gag}$ of Mo-MuLV contains an additional polypeptide chain of unknown function at the amino termi-

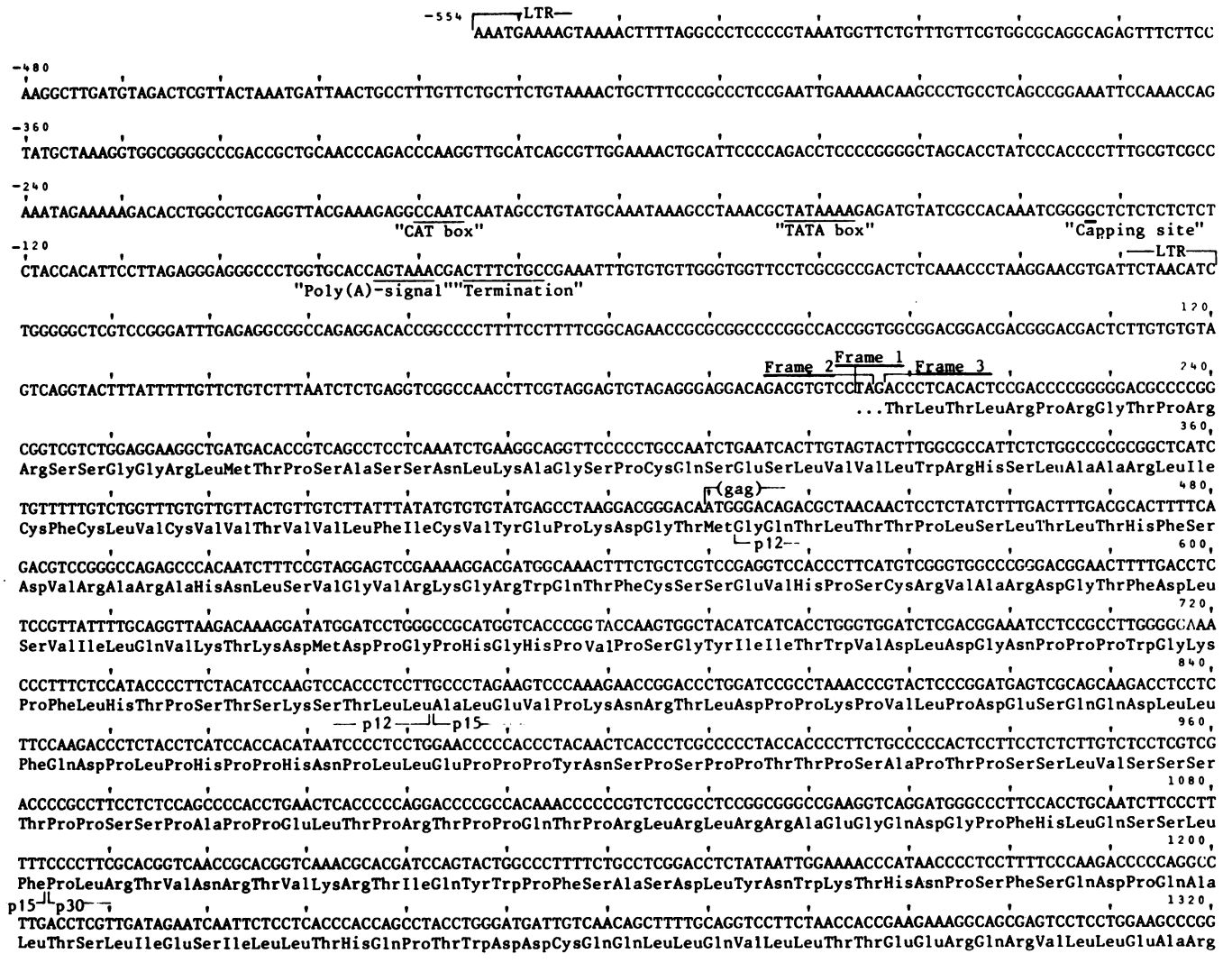
-554
LTR
AAATGAAAAGTAAAACTTTTAGGCCCTCCCCGTAAATGGTTCTGTTTGTTCGTGGCGCAGGCAGAGTTTCTTCC
-480
AAGGCTTGATGTAGACTCGTTACTAAATGATTAACTGCCTTTGTTCTGCTTCTGTAAAACTGCTTTCCCGCCCTCCGAATTGAAAAACAAGCCCTGCCTCAGCCGGAAATTCCAAACCAG
-360
TATGCTAAAGGTGGCGGGGCCCGACCGCTGCAACCCAGACCCAAGGTTGCATCAGCGTTGGAAAACTGCATTCCCCAGACCTCCCCGGGGCTAGCACCTATCCCACCCCTTTGCGTCGCC
-240
AAATAGAAAAAGACACCTGGCCTCGAGGTTACGAAAGAGGCCAATCAATAGCCTGTATGCAAATAAAGCCTAAACGCTATAAAAGAGATGTATCGCCACAAATCGGGGCTCTCTCTCTCT
"CAT box" "TATA box" "Capping site"
-120
LTR
CTACCACATTCCTTAGAGGGAGGGCCCTGGTGCACCAGTAAACGACTTTCTGCCGAAATTTGTGTGTTGGGTGGTTCCTCGCGCCGACTCTCAAACCCTAAGGAACGTGATTCTAACATC
"Poly(A)-signal" "Termination"
120
TGGGGGCTCGTCCGGGATTTGAGAGGCGGCCAGAGGACACCGGCCCCTTTTCCTTTTCGGCAGAACCGCGCGGCCCCGGCCACCGGTGGCGGACGGACGACGGGACGACTCTTGTGTGTA
Frame 2 Frame 1 Frame 3
240
GTCAGGTACTTTATTTTTGTTCTGTCTTTAATCTCTGAGGTCGGCCAACCTTCGTAGGAGTGTAGAGGGAGGACAGACGTGTCCTAGACCCTCACACTCCGACCCCGGGGGACGCCCCGG
...ThrLeuThrLeuArgProArgGlyThrProArg
360
CGGTCGTCTGGAGGAAGGCTGATGACACCGTCAGCCTCCTCAAATCTGAAGGCAGGTTCCCCCTGCCAATCTGAATCACTTGTAGTACTTTGGCGCCATTCTCTGGCCGCGCGGCTCATC
ArgSerSerGlyGlyArgLeuMetThrProSerAlaSerSerAsnLeuLysAlaGlySerProCysGlnSerGluSerLeuValValLeuTrpArgHisSerLeuAlaAlaArgLeuIle
(gag)
480
TGTTTTTGTCTGGTTTGTGTTGTTACTGTTGTCTTATTTATATGTGTGTATGAGCCTAAGGACGGGACAATGGGACAGACGCTAACAACTCCTCTATCTTTGACTTTGACGCACTTTTCA
CysPheCysLeuValCysValValThrValValLeuPheIleCysValTyrGluProLysAspGlyThrMetGlyGlnThrLeuThrThrProLeuSerLeuThrLeuThrHisPheSer
p12
600
GACGTCCGGGCCAGAGCCCACAATCTTTCCGTAGGAGTCCGAAAAGGACGATGGCAAACTTTCTGCTCGTCCGAGGTCCACCCTTCATGTCGGGTGGCCCGGGACGGAACTTTTGACCTC
AspValArgAlaArgAlaHisAsnLeuSerValGlyValArgLysGlyArgTrpGlnThrPheCysSerSerGluValHisProSerCysArgValAlaArgAspGlyThrPheAspLeu
720
TCCGTTATTTTGCAGGTTAAGACAAAGGATATGGATCCTGGGCCGCATGGTCACCCGGTACCAAGTGGCTACATCATCACCTGGGTGGATCTCGACGGAAATCCTCCGCCTTGGGGCAAA
SerValIleLeuGlnValLysThrLysAspMetAspProGlyProHisGlyHisProValProSerGlyTyrIleIleThrTrpValAspLeuAspGlyAsnProProProTrpGlyLys
840
CCCTTTCTCCATACCCCTTCTACATCCAAGTCCACCCTCCTTGCCCTAGAAGTCCCAAAGAACCGGACCCTGGATCCGCCTAAACCCGTACTCCCGGATGAGTCGCAGCAAGACCTCCTC
ProPheLeuHisThrProSerThrSerLysSerThrLeuLeuAlaLeuGluValProLysAsnArgThrLeuAspProProLysProValLeuProAspGluSerGlnGlnAspLeuLeu
p12 p15
960
TTCCAAGACCCTCTACCTCATCCACCACATAATCCCCTCCTGGAACCCCCACCCTACAACTCACCCTCGCCCCCTACCACCCCTTCTGCCCCCACTCCTTCCTCTCTTGTCTCCTCGTCG
PheGlnAspProLeuProHisProProHisAsnProLeuLeuGluProProProTyrAsnSerProSerProProThrThrProSerAlaProThrProSerSerLeuValSerSerSer
1080
ACCCCGCCTTCCTCTCCAGCCCCACCTGAACTCACCCCCAGGACCCCGCCACAAACCCCCCGTCTCCGCCTCCGGCGGGCCGAAGGTCAGGATGGGCCCTTCCACCTGCAATCTTCCCTT
ThrProProSerSerProAlaProProGluLeuThrProArgThrProProGlnThrProArgLeuArgLeuArgArgAlaGluGlyGlnAspGlyProPheHisLeuGlnSerSerLeu
1200
TTTCCCCTTCGCACGGTCAACCGCACGGTCAAACGCACGATCCAGTACTGGCCCTTTTCTGCCTCGGACCTCTATAATTGGAAAACCCATAACCCCTCCTTTTCCCAAGACCCCCAGGCC
PheProLeuArgThrValAsnArgThrValLysArgThrIleGlnTyrTrpProPheSerAlaSerAspLeuTyrAsnTrpLysThrHisAsnProSerPheSerGlnAspProGlnAla
p15 p30
1320
TTGACCTCGTTGATAGAATCAATTCTCCTCACCCACCAGCCTACCTGGGATGATTGTCAACAGCTTTTGCAGGTCCTTCTAACCACCGAAGAAAGGCAGCGAGTCCTCCTGGAAGCCCGG
LeuThrSerLeuIleGluSerIleLeuLeuThrHisGlnProThrTrpAspAspCysGlnGlnLeuLeuGlnValLeuLeuThrThrGluGluArgGlnArgValLeuLeuGluAlaArg

FIG. 3. Nucleotide sequences from the LTR to the *gag-pol* region of M7. The organization of the DNA is permuted with respect to the linear form of the viral DNA. Polarity is identical to that of the viral genome, and the nucleotides are numbered from the G residue just downstream from the LTR. Designations of the three reading frames are shown. The amino acid sequence has been deduced from the base sequence in reading frame 1. Important structures for transcription of the viral RNA, a CAT box, a TATA box, a capping site, and polyadenylation [Poly(A)] and termination signals are indicated, as reported previously. The *gag* and *pol* areas and the four mature *gag* proteins, p12, p15, p30, and p10, are shown. The sequence ATAA (positions 1,780 through 1,783) indicates a possible target sequence for the LTR-derived recombination as stated previously (42). The sites of restriction enzymes having 6-bp recognition located downstream from the LTR are as follows. The numbers represent the position of the 5′-end nucleotide generated by digestion with each enzyme. *Aca*I, 325, 1125; *Acc*I, 1,582, 1,736; *Acy*I, 232, 482; *Ava*I, 2,054; *Avr*II, 2,230; *Asu*II, 1,575; *Bam*HI, 634, 793, 1,689; *Ban*II 5,495, 1,055, 2,039, 2,046, 2,057, 2,346; *Bgl*II, 2,299; *Cfr*I, 28, 345; *Hae*I, 1,197, 1,342, 1,500, 1,626, 2,368; *Hin*dII, 1,097, 1,257; *Hin*HI, 333, 2,114; *Hpa*I, 2,225; *Kpn*I, 2,266; *Pst*I, 1,706; *Sal*I, 958; *Sst*I, 2,319; *Xma*I, 225, 2,114; *Xho*II, 688, 2,325.

TABLE 1. Analysis of open reading frames in the 5'-distal region of M7 DNA[a]

Reading frame	Initiation codon	Termination codon		
	ATG	TAG (amber)	TGA (opal)	TAA (ochre)
1	262	175		
	429	205		
	631	2,026		
	1,551			
	1,579			
	1,620			
	2,197			
	2,242			
2	410	−107	20	149
	647	119	260	443
	818	323	263	1,280
	1,052	512	287	1,508
	1,250	767	461	1,517
	1,325	1,214	467	1,651
	1,354	1,370	1,201	1,718
	1,409	1,630	1,211	1,718
	1,352	1,645	1,856	1,826
	1,700	1,682	2,132	2,066
	2,366	1,715	2,246	2,309
	2,405	1,853	2,387	
	2,450	1,907		
		2,009		
		2,078		
		2,231		
		2,357		
3	402	183	−13	−82
	531	1,425	312	−22
	567	1,773	411	−7
	2,205	1,974	594	417
	2,274		819	618
			987	801
			1,251	870
			1,410	1,155
			1,583	1,170
			1,860	1,743
			2,013	1,902
			2,271	1,911
			2,406	2,211
			2,421	2,226

[a] All of the initiation and termination codons from the capping site to the A residue at position 2,469 are listed. Numbers indicate the position of the initial base of each codon. Designation of frames 1, 2, and 3 is shown in Fig. 3.

nus. By analogy, the region from position 262 to 429 in M7 DNA may encode amino acids in addition to the major *gag* precursor, Pr68, which may be eliminated during the maturation process of the *gag* protein. Alternatively, a splicing event between an acceptor-like sequence at position 415, CCTAAG, and a donor sequence at 122, TCAG/GTAC, may occur. If this is the case, the mRNA encoding the viral proteins might be different from the genomic RNA present in M7 virions.

Characteristics of M7 *gag* peptides as deduced from DNA sequence analysis. The polypeptide structure of the mature M7 *gag* precursor is NH_2-p12-p15-p30-p10-COOH (2, 25, 39). The NH_2-terminal amino acid of p12 is modified in vivo (8), and its identity has not yet been determined. On the other hand, the *gag* gene of Mo-MuLV contiguously encodes the protein NH_2-p15-p12-p30-p10-COOH. As shown below in Fig. 4, the nucleotide sequence in the p12 region of M7 is strongly homologous to that in the p15 region of Mo-MuLV. The initial amino acid of the p15 region is glycine. From the sequence data shown in Fig. 3, the first amino acid of the mature M7 p12 is glycine. The molecular weight of the M7 p12, as estimated from its predicted amino acid sequence, is 11,900.

The second *gag* polypeptide, p15, is a phosphoprotein and may be associated with the genomic RNA in the M7 virion (26, 34). Peptide p12 of Mo-MuLV is also a phosphoprotein, with a molecular weight of 9,200. Interestingly, the estimated molecular weight of M7 p15 is 11,600. I suggest that the p15 of M7 is either linked with chemical groups which give the protein a more positive charge or that the protein has an unusual configuration, thus accounting for this apparent discrepancy.

The major structural protein of the virion core is p30. Several groups have reported the base and amino acid sequences in the p30 region of M7 (7, 8, 24, 25), and the present results are consistent with these reports.

The fourth peptide derived from the M7 *gag* precursor is p10. The carboxy-terminal amino acid of this peptide is reported to be leucine (8), and the codon at either position 2,002 or 2,008 in the DNA sequence of Fig. 3 could encode this residue. Based on similarities of the amino acid sequences observed between the M7 and Mo-MuLV *gag* polyproteins (Fig. 4), I propose that the carboxy-terminal leucine of the M7 p10 is derived from the codon at position 2,002 and is followed by a threonine-leucine sequence. Several other residues are apparently lost from the carboxy end of the original *gag* precursor molecule.

Thus, based on the observations reported here, the entire M7 *gag* polyprotein gene is 1,572 bases in length. The four M7 *gag* polypeptides are coded in a contiguous manner from a single open reading frame.

I have analyzed the content of basic or hydrophobic amino acids in each of four M7 polypeptides. The amino-terminal peptide of retroviruses is generally very hydrophobic (1, 10). In fact, a stretch of hydrophobic amino acids was detected in the M7 p12 region beginning at

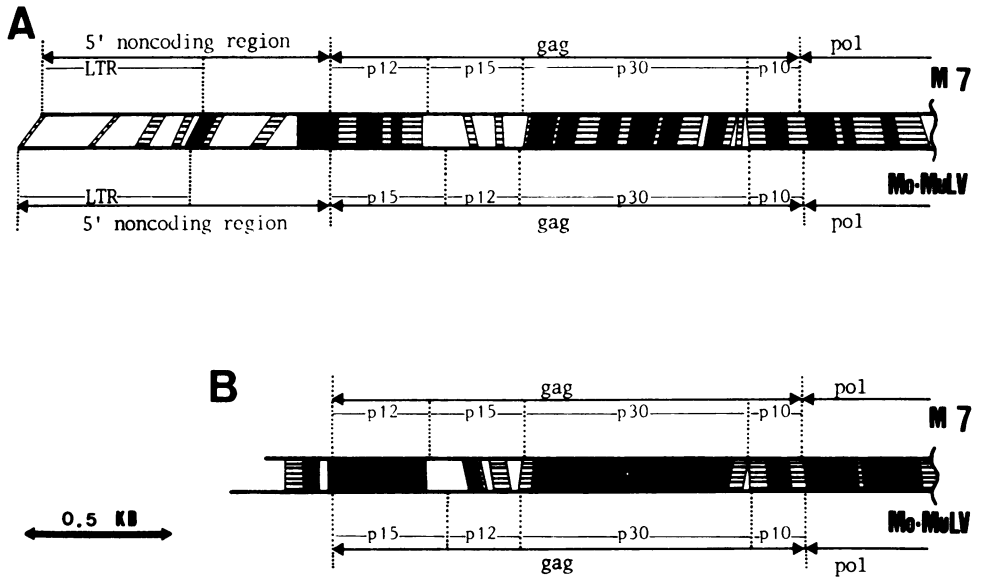

FIG. 4. Homologies of nucleotide and amino acid sequences between M7 and Mo-MuLV. Nucleotide (A) and amino acid (B) sequences in the 5'-distal region of the M7 DNA were compared with those of Mo-MuLV by computer analysis. The homologous regions in these two viruses are illustrated. Sequence commonalities were calculated within each stretch of 24 bases (A) or of eight amino acids (B). (A) Regions of base sequences with 75% or greater commonality, solid boxes; those with 58 to 74% commonality, horizontally lined boxes; those with 57% or less commonality, open boxes. (B) Regions of amino acid sequences with 63% or greater commonality, solid boxes; those with 38 to 62% commonality, horizontally lined boxes; those with 37% or less commonality, open boxes.

position 703 (Fig. 3). These amino acids may assume a salient position on the external surface of the protein molecule where they may be more readily associated with the envelope of the virion. The second *gag* protein of mammalian retroviruses, corresponding to p15 of of M7 virus, is moderately acidic, whereas the fourth polypeptide, corresponding to p10 of M7 virus, is known to be highly basic (10, 14, 39). The present results are consistent with these findings. Based on the content of hydrophobic amino acids, the p15 of M7 and the p12 of Mo-MuLV appear to be the most hydrophobic *gag* proteins. However, these peptides do not exhibit hydrophobicity (10). Conceivably, the tertiary structure of these peptides is responsible for this property.

Coding region for M7 reverse transcriptase. The *pol* gene of retroviruses is usually located directly downstream from the *gag* gene (6). I found a long open reading frame in M7 DNA following an amber termination codon at position 2,026. This appears to constitute part of the coding region of the *pol* gene of this virus. Significant amino acid and base homologies in the NH_2 termini of the reverse transcriptases (RTs) of M7 and Mo-MuLV were observed (see Fig. 4), and these homologies also suggest a reading frame for *pol* starting at position 2.029. Although an ATG codon is located at position 2,197, it apparently does not code for the first amino acid of the RT molecule. The Mo-MuLV RT is initially synthesized as a large precursor protein, Pr180 (28, 46), encoded by the *gag-pol* region. Only one amber termination codon interrupts the *gag* and *pol* genes of Mo-MuLV, and a read-through mechanism bypassing this UAG codon appears to play an important role in the synthesis of Mo-MuLV RT. I suggest that a similar read-through mechanism is involved in the synthesis of the corresponding M7 enzyme.

Nucleotide and amino acid sequence homologies between M7 and Mo-MuLV. Mo-MuLV is a prototype for mammalian type C retroviruses, and its complete base sequence has been reported previously (37). Using computer analysis, I examined homologies in the base and amino acid sequences between these two viruses (Fig. 4). Base sequences in the regions of p12, p30, p10, and the 5'-distal portion of the *pol* gene are 50, 59, 60, and 59% homologous, respectively. A similar degree of homology (50 to 65%) was observed between the corresponding amino acid sequences of these viruses. I suggest that the 5' coding regions of retroviruses have been conserved throughout a long evolutionary period. Bonner et al. (4) recently obtained similar results

for the other retroviral genomes. Conservation of base and amino acid sequences in the proximal p30 region (>80%) is not surprising, since these major core proteins are known to be serologically related to each other (10, 38).

Amino acids at the amino and carboxy termini for each of the four *gag* polypeptides of M7 virus are similar to those of Mo-MuLV except at the junction of p12-p15 (Fig. 4). These sequences may be highly conserved because they are essential for cleavage of the *gag* precursor molecule. Amino acid and base sequences at the junction of p12-p15 in M7 virus are completely different from those at the corresponding junction (p15-p12) in Mo-MuLV. Moreover, the p15 region of M7 does not have any significant homologies to the p12 region of Mo-MuLV. Barbacid et al. (2) have suggested that, in the M7 virus, the cleavage site between the p12 and p15 polypeptides has shifted toward the 5′ terminus of the viral genome. The data from this report support this assumption. I suggest that if, as some studies indicate (34), the p15 phosphoprotein of M7 acts by binding to its homologous RNA genome, then structural variations observed among retroviral phosphoproteins may reflect a requirement for retaining a degree of binding specificity to their homologous viral genome.

As shown in Fig. 4, the 5′ noncoding region of M7 virus, comprising about 150 bases immediately upstream from the p12-coding region, shows homologies of 55 and 35% in base sequence and amino acid sequence, respectively, to Mo-LuLV. This region could encode some viral proteins which have not as yet been identified, as suggested by Mo-MuLV (37). A potential initiation codon (ATG) for such a protein can be found 171 bases upstream from the p12 region of M7.

ACKNOWLEDGMENTS

I thank Hisao Uchida of the Institute of Medical Sciences, Tokyo University, Tokyo, Japan, for his help on the computer analysis. I also thank Robert H. Bassin of the National Institutes of Health, Bethesda, Md., for reading the manuscript.

This work was supported in part by research grants to T.T. and to Toshiya Takano of the Department of Microbiology, Keio University School of Medicine, Tokyo, Japan, from the Japanese Ministry of Education, Science and Culture, the Naito Foundation, and the Waksman Foundation of Japan.

LITERATURE CITED

1. **Barbacid, M., and S. A. Aaronson.** 1978. Membrane properties of the *gag* gene-coded p15 protein of mouse type-C RNA tumor viruses. J. Biol. Chem. **253**:1408–1414.
2. **Barbacid, M., J. R. Stephenson, and S. A. Aaronson.** 1977. Evolutionary relationships between *gag* gene-coded proteins of murine and primate endogenous type C viruses. Cell **10**:641–648.
3. **Bishop, J. M.** 1978. Retroviruses. Annu. Rev. Biochem. **47**:35–88.
4. **Bonner, T. I., C. O'Connell, and M. Cohen.** 1982. Cloned endogenous retroviral sequences from human DNA. Proc. Natl. Acad. Sci. U.S.A. **79**:4709–4713.
5. **Breathnach, R., C. Benoist, K. O'Hare, F. Gannon, and P. Chambon.** 1978. Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. Proc. Natl. Acad. Sci. U.S.A. **75**:4853–4857.
6. **Coffin, J. M.** 1982. Structure of the retroviral genome, p. 261–368. *In* R. A. Weiss, N. M. Teich, H. E. Varmus, and J. M. Coffin (ed.), RNA tumor viruses. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
7. **Cohen, M., A. Rein, R. M. Stephens, C. O'Connell, R. V. Gilden, M. Shure, M. O. Nicolson, R. M. McAllister, and N. Davidson.** 1981. Baboon endogenous virus genome: molecular cloning and structural characterization on non-defective viral genomes from DNA of a baboon cell strain. Proc. Natl. Acad. Sci. U.S.A. **78**:5207–5211.
8. **Copeland, T. D., L. E. Henderson, E. S. Vanlaningham-Miller, J. R. Stephenson, G. W. Smythers, and S. Oroszlan.** 1981. Amino- and carboxyl-terminal sequences of proteins coded by *gag* gene of endogenous baboon and cat type C viruses. Virology **109**:13–24.
9. **Darlix, J.-L., M. Zuker, and P.-F. Spahr.** 1982. Structure-function relationship of Rous sarcoma virus leader RNA. Nucleic Acids Res. **10**:5183–5196.
10. **Dickson, C., R. N. Eisenman, H. Fan, E. Hunter, and N. M. Teich.** 1982. Protein biosynthesis and assembly, p. 513–648. *In* R. A. Weiss, N. M. Teich, H. E. Varmus, and J. M. Coffin (ed.), RNA tumor viruses. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
11. **Edwards, S. A., and H. Fan.** 1979. *gag*-Related polyproteins of Moloney murine leukemia virus: evidence for independent synthesis of glycosylated and unglycosylated forms. J. Virol. **30**:551–563.
12. **Eisenman, R. N., and V. M. Vogt.** 1978. The biosynthesis of oncovirus proteins. Biochim. Biophys. Acta **473**:187–239.
13. **Fan, H., and I. M. Verma.** 1978. Size analysis and relationship of murine leukemia virus-specific mRNA's: evidence for transposition of sequences during synthesis and processing of subgenomic mRNA. J. Virol. **26**:468–478.
14. **Fleissner, E., and E. Tress.** 1973. Isolation of a ribonucleoprotein structure from oncornaviruses. J. Virol. **12**:1612–1615.
15. **Hackett, P. H., R. Swanstrom, H. E. Varmus, and J. M. Bishop.** 1982. The leader sequence of the subgenomic mRNA's of Rous sarcoma virus is approximately 390 nucleotides. J. Virol. **41**:527–534.
16. **Hayward, W. S.** 1977. Size and genetic content of viral RNAs in avian oncovirus-infected cells. J. Virol. **24**:47–63.
17. **Hughes, S. H., P. R. Shank, D. H. Spector, H.-J. King, J. M. Bishop, H. E. Varmus, P. K. Vogt, and M. L. Breitman.** 1978. Proviruses of avian sarcoma virus are terminally redundant, co-extensive with unintegrated linear DNA and integrated at many sites. Cell **15**:1397–1410.
18. **Jamjoom, G. A., R. B. Naso, and R. B. Arlinghaus.** 1977. Further characterization of intracellular precursor polyproteins of Rauscher leukemia virus. Virology **78**:11–34.
19. **Kitamura, N., B. L. Semler, P. G. Rothberg, G. R. Larsen, C. J. Adler, A. J. Dorner, E. A. Emini, R. Hanecak, J. J. Lee, S. Van der Werf, C. W. Anderson, and E. Wimmer.** 1981. Primary structure, gene organization and polypeptide expression of poliovirus RNA. Nature (London) **291**:547–553.
20. **Kozak, M.** 1978. How do eukaryotic ribosomes select initiation regions in messenger RNA? Cell **15**:1109–1123.
21. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **55**:499–560.
22. **Murphy, E. C., Jr., D. Campos, III, and R. B. Arlinghaus.** 1979. Cell-free synthesis of Rauscher murine leukemia

virus "gag" and "env" gene products from separate cellular mRNA species. Virology 93:293–302.
23. **Noda, M., M. Wagatsuma, T. Tamura, T. Takano, and K.-I. Matsubara.** 1981. Structure of the baboon endogenous virus genome: cloning of circular virus DNA in bacteriophage λ. Nucleic Acids Res. 9:2173–2185.
24. **Oroszlan, S., T. Copeland, M. R. Summers, G. Smythers, and R. V. Gilden.** 1975. Amino acid sequence homology of mammalian type C RNA virus major internal proteins. J. Biol. Chem. 250:6232–6239.
25. **Oroszlan, S., M. Summers, and R. V. Gilden.** 1975. Amino-terminal sequence of baboon type C p30. Virology 64:581–583.
26. **Pal, B. K., and P. Roy-Burman.** 1975. Phosphoproteins: structural components of oncornaviruses. J. Virol. 15:540–549.
27. **Peters, G., F. Harada, J. E. Dahlberg, A. Panet, W. A. Haseltine, and D. Baltimore.** 1977. Low-molecular-weight RNAs of Moloney murine leukemia virus: identification of the primer for RNA-directed DNA synthesis. J. Virol. 21:1031–1041.
28. **Philipson, L., P. Andersson, U. Olshevsky, R. Weinberg, D. Baltimore, and R. Gesteland.** 1978. Translation of MuLV and MSV RNAs in nuclease-treated reticulocyte extracts: enhancement of the *gag-pol* polypeptide with yeast suppressor tRNA. Cell 13:189–199.
29. **Reddy, E. P., M. J. Smith, and S. A. Aaronson.** 1981. Complete nucleotide sequence and organization of the Moloney murine sarcoma virus genome. Science 214:445–450.
30. **Rothenberg, E., D. J. Donoghue, and D. Baltimore.** 1978. Analysis of a 5′ leader sequence on murine leukemia virus 21 S RNA: heteroduplex mapping with long reverse transcriptase products. Cell 13:435–451.
31. **Sabran, J. L., T. W. Hsu, C. Yeater, A. Kaji, W. S. Mason, and J. M. Taylor.** 1979. Analysis of integrated avian RNA tumor virus DNA in transformed chicken, duck, and quail fibroblasts. J. Virol. 29:170–178.
32. **Sawyer, R. C., and J. E. Dahlberg.** 1973. Small RNAs of Rous sarcoma virus: characterization by two-dimensional polyacrylamide gel electrophoresis and fingerprint analysis. J. Virol. 12:1226–1237.
33. **Schultz, A. M., and S. Oroszlan.** 1978. Murine leukemia virus *gag* polyproteins: the peptide chain unique to Pr80 is located at the amino terminus. Virology 91:481–486.
34. **Sen, A., C. J. Sherr, and G. J. Todaro.** 1978. Endogenous feline (RD-114) and baboon type C viruses have related specific RNA-binding proteins and genome binding sites. Virology 84:99–107.
35. **Shank, P. R., S. H. Hughes, H. S. Kung, J. E. Majors, N. Quintrell, R. V. Guntaka, J. M. Bishop, and H. E. Varmus.** 1978. Mapping unintegrated avian sarcoma virus DNA: termini of linear DNA bear 300 nucleotides present once or twice in two species of circular DNA. Cell 15:1383–1395.
36. **Sharp, P. A.** 1981. Speculations on RNA splicing. Cell 23:643–646.
37. **Shinnick, T. M., R. A. Lerner, and J. G. Sutcliffe.** 1981. Nucleotide sequence of Moloney murine leukemia virus. Nature (London) 293:543–548.
38. **Stephenson, J. R.** 1980. Tye C virus structural and transformation-specific proteins, p. 245–297. *In* J. R. Stephenson (ed.), Molecular biology of RNA tumor viruses. Academic Press, Inc., N.Y.
39. **Stephenson, J. R., R. K. Reynolds, S. G. Devare, and F. H. Reynolds.** 1977. Biochemical and immunological properties of *gag* gene-coded structural proteins of endogenous type C RNA tumor viruses of diverse mammalian species. J. Biol. Chem. 252:7818–7825.
40. **Swanstrom, R., H. E. Varmus, and J. M. Bishop.** 1982. Nucleotide sequence of the 5′ noncoding region and part of the *gag* gene of Rous sarcoma virus. J. Virol. 41:535–541.
41. **Tamura, T., M. Noda, and T. Takano.** 1981. Structure of the baboon endogenous virus genome: Nucleotide sequences of the long terminal repeat. Nucleic Acids Res. 9:6615–6626.
42. **Tamura, T., and T. Takano.** 1982. Long terminal repeat (LTR)-derived recombination of retroviral DNA: sequence analyses of an aberrant clone of baboon endogenous virus DNA which carries an inversion from the LTR to the *gag* region. Nucleic Acids Res. 10:641–648.
43. **Taylor, J. M.** 1977. An analysis of the role of tRNA species as primers for the transcription into DNA of RNA tumor virus genomes. Biochim. Biophys. Acta 475:57–71.
44. **Van Beveren, C., J. G. Goddard, A. Berns, and I. M. Verma.** 1980. Structure of Moloney murine leukemia viral DNA: nucleotide sequence of the 5′ long terminal repeat and adjacent cellular sequences. Proc. Natl. Acad. Sci. U.S.A. 77:3307–3311.
45. **Van Beveren, C., F. van Straaten, J. A. Galleshaw, and I. M. Verma.** 1981. Nucleotide sequence of the genome of a murine sarcoma virus. Cell 27:97–108.
46. **Verma, I. M.** 1977. The reverse transcriptase. Biochim. Biophys. Acta 473:1–38.
47. **Vogt, V. M., and R. Eisenman.** 1973. Identification of a large polypeptide precursor of avian oncornavirus protein. Proc. Natl. Acad. Sci. U.S.A. 70:1734–1738.
48. **Weiss, S. R., H. E. Varmus, and J. M. Bishop.** 1977. The size and genetic composition of virus-specific RNAs in cytoplasm of cells producing avian sarcoma-leukemia viruses. Cell 12:983–992.