JOURNAL OF VIROLOGY, Nov. 1986, p. 436–449
0022-538X/86/110436-14$02.00/0

Vol. 60, No. 2

Conserved TAAATG Sequence at the Transcriptional and Translational Initiation Sites of Vaccinia Virus Late Genes Deduced by Structural and Functional Analysis of the *Hin*dIII H Genome Fragment

JOHANNES L. ROSEL,† PATRICIA L. EARL, JERRY P. WEIR, AND BERNARD MOSS*

Laboratory of Viral Diseases, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland 20892

Received 28 April 1986/Accepted 13 July 1986

The sequence of the 8,600-base-pair *Hin*dIII H fragment, located at the center of the vaccinia virus genome, was determined to analyze several late genes. Seven major complete open reading frames (ORFs) and two that started from or continued into adjacent DNA segments were identified. ORFs were closely spaced and present on both DNA strands. Some adjacent ORFs had oppositely oriented overlapping termination codons or contiguous stop and start codons. Nucleotide compositional analysis indicated that the A-T frequency was consistently lowest in the first codon position. The sizes of the polypeptides predicted from the DNA sequence were compared with those determined by polyacrylamide gel electrophoresis of cell-free translation products of mRNAs selected by hybridization to cloned single-stranded DNA segments or synthesized in vitro by bacteriophage T7 RNA polymerase. Six transcripts that initiated within the *Hin*dIII H DNA fragment were detected, and of these, four were synthesized only at late times, one was synthesized only early, and one was synthesized early and late. The sites on the genome corresponding to the 5′ ends of the transcripts were located by high-resolution nuclease S1 analysis. For late genes, the transcriptional and translational initiation sites mapped within a few nucleotides of each other, and in each case the sequence TAAATGG occurred at the start of the ORF. The extremely short leader and the absence of A or G in the −3 position, relative to the first nucleotide of the initiation codon, distinguishes the majority of vaccinia virus late genes from eucaryotic and vaccinia virus early genes.

Vaccinia virus, a member of the poxvirus family, has a 185,000-base-pair (bp) double-stranded DNA genome, uses its own transcriptional apparatus, and replicates in the cytoplasm of infected cells (25). The genes, which may number between 150 and 200, are regulated in a temporal fashion. The early genes are transcribed by the RNA polymerase within the vaccinia virus core, whereas expression of late genes is delayed until virus uncoating and DNA replication occur. Clusters of early and late genes are interspersed throughout the DNA molecule (2), and the signals that regulate their expression are located just upstream of the transcriptional initiation sites (3, 8, 21, 34). Examination of the region preceding several early RNA start sites revealed extremely A-T-rich sequences that lack highly conserved procaryotic or eucaryotic consensus signals (32, 33). The RNA polymerase and associated factors extracted from virus particles specifically recognize early genes and distinguish them from late genes (14, 27). Too few late genes have been analyzed in detail (3, 28, 34) to develop a consensus regulatory sequence.

The present study was undertaken to learn more about the organization and regulation of late genes. The *Hin*dIII H fragment was chosen for analysis because previous reports indicated that several late proteins are encoded within this region of the vaccinia virus genome (1, 2, 22). We have completed the sequence of this 8,600-bp DNA fragment, identified the open reading frames (ORFs), and characterized some of the gene products. Most significantly, the leader between the 5′ end of the late mRNAs and their translational initiation codons appeared to be extremely short and the DNA at that junction contained a conserved sequence TAAATGG.

MATERIALS AND METHODS

Virus and cells. Vaccinia virus (strain WR) was grown in HeLa cell suspension cultures that were maintained in Eagle medium containing 5% horse serum.

DNA purification and sequencing. Plasmid DNA was purified as described by Birnboim and Doly (4). DNA fragments were isolated by electroblotting onto DEAE paper (36) and subcloned into mp18 or mp19 derivatives of bacteriophage M13 (23). Exonuclease III digestions were carried out (16), and sequencing was performed as described by Sanger and co-workers (29). All sequences were determined on both DNA strands and were completely overlapped with the aid of additional specific oligonucleotide primers.

RNA analysis. HeLa cells were infected at a multiplicity of 30 PFU of purified vaccinia virus per cell for 4 h in the presence of 100 μg of cycloheximide per ml to obtain early RNA or 6 h in the absence of the drug to obtain late RNA. Cytoplasmic RNA was isolated by CsCl centrifugation as described before (10). RNA was hybridized to asymmetrically end-labeled DNA fragments, and single-stranded DNA was digested with 580 U of S1 nuclease or 240 U of mung bean nuclease at 25°C for 1 h (34). The nuclease-resistant DNA was analyzed by electrophoresis on agarose or polyacrylamide sequencing gels.

* Corresponding author.
† Present address: CIBA-GEIGY, Basel, Switzerland.

In vitro translation. Purified cytoplasmic RNA was hybridized to plasmid or single-stranded phage DNA immobilized on nitrocellulose filters (10). After washing, the specifically bound RNA was eluted, ethanol precipitated with carrier tRNA, and then translated in a micrococcal nuclease-treated reticulocyte lysate containing [^{35}S]methionine. Radioactively labeled polypeptides were analyzed by polyacrylamide gel electrophoresis.

In vitro transcription. DNA fragments were cloned by using a Bluescribe M13+ plasmid vector (Vector Cloning Systems). The recommendations of the manufacturer were followed for synthesis of RNA except that T7 RNA polymerase and template were incubated with 1 μM m^{7}GpppG and m^{7}GpppA for 10 min prior to addition of ribonucleoside triphosphates. The reaction mixture was DNase treated, phenol extracted, and ethanol precipitated prior to translation.

Computer analysis of DNA sequences. DNA sequence data were managed and analyzed by using the Microgenie (Beckman Instruments, Inc.) program on an IBM XT personal computer. Similarity searches of the National Biomedical Research Foundation protein library were carried out by using the FASTP program written by Lipman and Pearson (20). Initial screening was done with a ktup of 2.

RESULTS

Nucleotide sequence of the *Hin*dIII H fragment. The restriction enzyme *Hin*dIII cleaves the 185,000-bp genome of vaccinia virus into 15 fragments ranging in size from 1,500 to 45,000 bp (12, 37). The *Hin*dIII H fragment is about 8,600 bp long and is located at the center of the genome in a region with several late genes (1, 2, 22). The sequence at the extreme left end of the *Hin*dIII H fragment was previously determined and was shown to contain the distal half of an early gene encoding the RNA polymerase large subunit (7). Additional overlapping subclones of the *Hin*dIII H fragment in M13 vectors were generated by exonuclease III digestion and sequenced by the dideoxynucleotide chain termination method. Both strands were sequenced completely, and additional primary DNA clones were obtained in regions of particular interest to exclude errors arising from mutations or deletions during cloning. The sequence of the entire *Hin*dIII H fragment is shown in Fig. 1.

Analysis of ORFs. The nucleotide sequence was translated in both directions in all possible phases to locate ORFs. There are eight major ORFs that begin with ATG in the *Hin*dIII H fragment and one that begins with an ATG in the adjacent *Hin*dIII J fragment and continues into H (Fig. 1). Of these, five were directed rightward and four were directed leftward. Major ORFs were named by using the letter designating the *Hin*dIII fragment in which they originated and then numbered successively from left to right. The reading direction is indicated by adding an L (left) or R (right) after the number. Thus, an ORF map of *Hin*dIII H can be abbreviated as J6R-H1L-H2R-H3L-H4L-H5L-H6R-H7R-H8R. The designation J6R for the extreme left ORF is based on previous analyses of that fragment (7, 26, 33).

The sizes of ORFs H1 to H7 range from 171 (H1) to 575 (H5) codons. H8 has 129 codons within the *Hin*dIII H fragment and an undetermined number within the adjacent *Hin*dIII D fragment. Based on the direction of the ORF, four types of junctions can be distinguished: R–L (→ ←), L–R (← →), R–R (→ →), and L–L (← ←). In general, the ORFs are very closely spaced. For the R–L junction of J6R and H1L, the stop codons of the two ORFs actually overlap for two of the three nucleotides (NT) and for H2R and H3L the top codons are separated by only two NT. There are also two L–R junctions. In the case of H1L and H2R, the ATGs are separated by 13 NT and for H5L and H6R they are separated by 184 NT. The two L–L junctions occur successively, and the gaps between the stop and start codons of H5L and H4L and H4L and H3L are 120 and 0 NT. The two R–R junctions also occur successively, and the gaps between the stop and start codons of H6R and H7R and H7R and H8R are 0 and 36 NT.

Since the vaccinia virus genome has a rather high A+T content, we considered that the three codon positions might exhibit a characteristic pattern of nucleotides. Such a pattern might be useful in distinguishing true coding sequences from chance ORFs. An analysis (Table 1) revealed that the A-T frequency was consistently lowest in the first codon position.

It was also of interest to determine the nucleotide frequencies around the ATG that starts each ORF. In eucaryotic mRNAs, there is a highly conserved pattern with an A or G residue almost always present at the 3′ position relative to the A of the ATG and a G usually present at the +4 position (18). Consistent with this, a G was found in the +4 position of six of seven ORFs in the *Hin*dIII H fragment. In the −3 position, however, a pyrimidine was present in six of seven cases and in five of those the sequence was TAAATGG. Evidence that the latter sequence correlates with late genes will be presented.

Computer analyses of ORFs were performed to predict amino acid compositions (Table 2), secondary structure, hydrophobicity, and similarity to previous entries in the National Biomedical Research Foundation protein library. The sequence of the large RNA polymerase subunit, ORF J6R, and its homology to the B′ subunit of *Escherichia coli* and to the large subunit of yeast RNA polymerases II and III have been reported previously (7) and will not be commented on further here.

H1L potentially encodes a basic 19.7-kilodalton (kDa) polypeptide lacking tryptophan. The presence of a string of 14 hydrophobic amino acids near the carboxyl terminus would make it interesting to find out whether the protein is membrane associated. Several proteins in the National Biomedical Research Foundation library showed modest similarities to H1L over short regions but none had obvious biological significance.

H2R could encode a 21.5-kDa basic polypeptide. No striking similarities to proteins in the data bank were noted, although the highest optimized score of 71 (20) was with the beta-chain precursor of the T-cell receptor.

H3L codes for an acidic 37.5-kDa polypeptide that contains two hydrophobic stretches of 14 and 16 amino acids near the carboxyl terminus. (Data to be shown later indicate the presence of a second RNA start site within the ORF, making it possible that two proteins with overlapping amino acid sequences are expressed from this ORF.) The highest degree of relatedness was a 22.2% identity in a 102-amino-acid overlap with the hydrophobic region of the spike glycoprotein precursor of vesicular stomatitis virus.

H4L and H5L could code for neutral 21.2 and 67.6-kDa polypeptides, respectively. Both ORFs only had short or modest matches with proteins in the data bank.

The 22.3-kDa polypeptide predicted by the H6R ORF is slightly acidic but with a concentration of basic amino acids near the carboxyl terminus, very hydrophilic, and relatively deficient in aromatic amino acids. Evidence for an unusual electrophoretic mobility on sodium dodecyl sulfate-

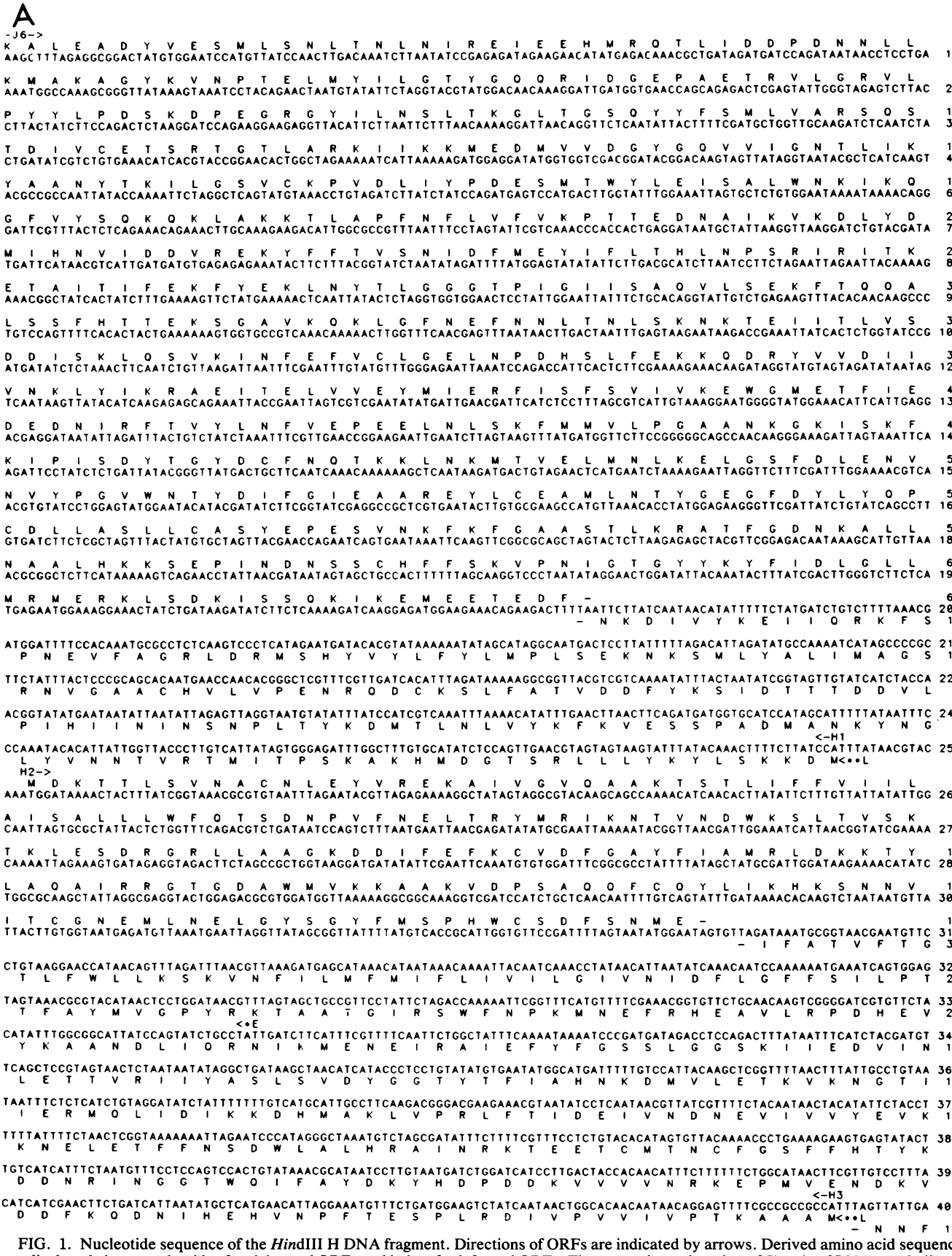

FIG. 1. Nucleotide sequence of the *Hin*dIII H DNA fragment. Directions of ORFs are indicated by arrows. Derived amino acid sequences are displayed above nucleotides for rightward ORFs and below for leftward ORFs. The approximate location of 5′ ends of RNAs are indicated by an asterisk, with an arrow for direction of transcription and E or L for early or late expression.

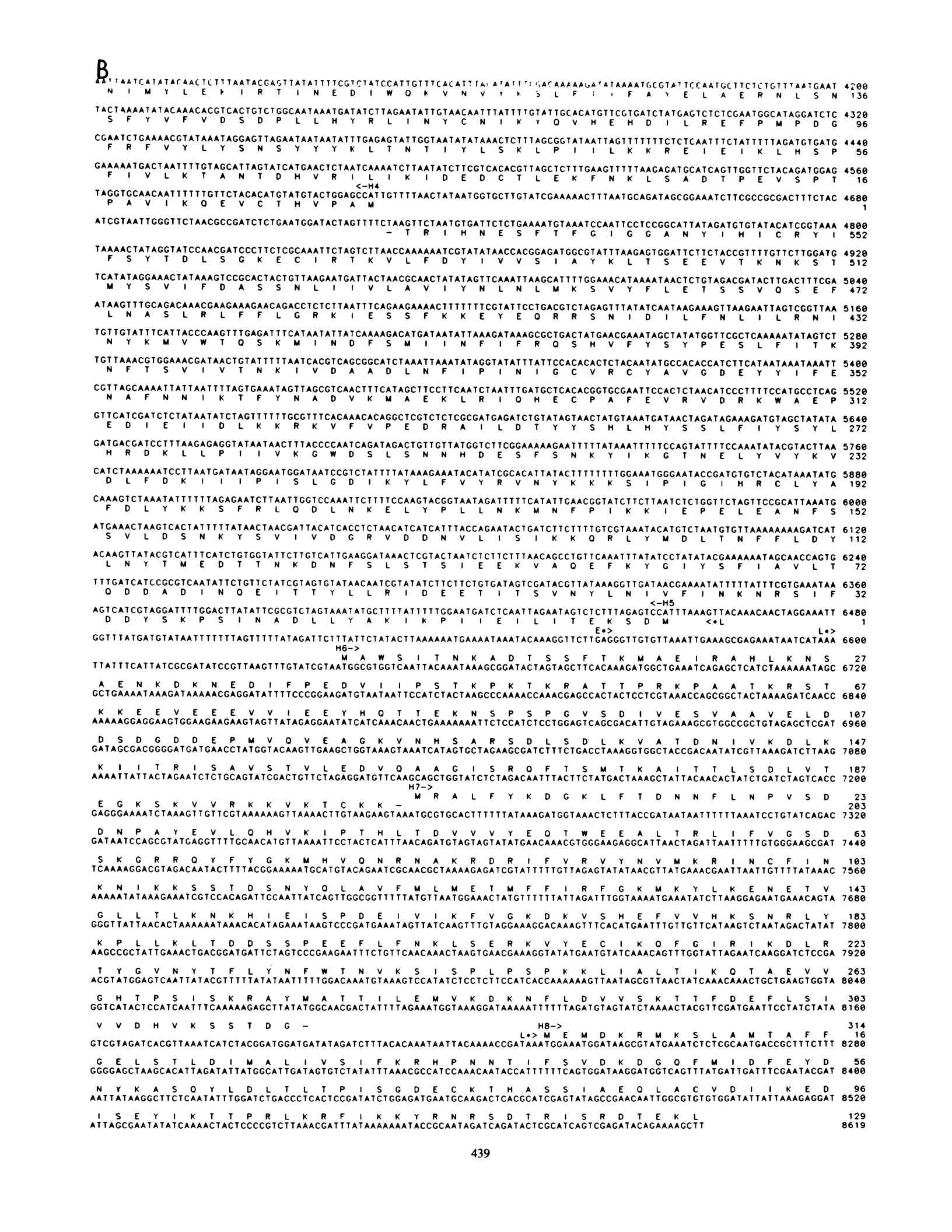

TABLE 1. Correlation of A-T frequency with codon position

Codon Position	% A + T								
	H1	H2	H3	H4	H5	H6	H7	H8	H av
1	63.0	60.0	55.7	62.6	64.1	50.0	59.4	58.2	59.1
2	69.2	63.1	70.8	74.3	73.5	60.8	70.8	65.9	68.4
3	68.7	69.5	68.0	74.9	68.3	73.5	75.6	64.4	70.4
1+2+3	67.3	64.3	64.8	70.6	68.6	61.4	68.5	62.8	66.0

containing polyacrylamide gels is presented below. Again, no impressive homology matches were noted.

H7R encodes a putative basic 36.7-kDa polypeptide with potential glycosylation sites at amino acids 142 and 230. A search of the National Biomedical Research Foundation protein library revealed a short region of similarity to ribonucleotide reductase of Epstein-Barr virus. Induction of ribonucleotide reductase in cells infected with vaccinia virus has been reported but the gene has not been mapped (30). Only 14 kDa of H8R is present in the *Hin*dIII H fragment and the remainder is presumably in the adjacent *Hin*dIII D fragment.

In vitro translation of mRNAs selected by hybridization to *Hin*dIII H DNA. In vitro translation experiments were carried out to determine whether mRNAs that could express the ORFs are synthesized in vaccinia virus-infected cells, to compare the polypeptide products with those predicted, and to ascertain the temporal regulation of the genes. By using single-stranded M13 subclones of the *Hin*dIII H fragment for hybridization selection, information regarding the location of translatable mRNAs relative to ORFs was obtained.

When early RNA that hybridized to the entire *Hin*dIII H fragment was translated in a nuclease-treated reticulocyte lysate, the most prominent band had an apparent molecular weight of about 40,000 (Fig. 2A). The light band of 147 kDa was previously shown to be the largest subunit of RNA polymerase and to be a product of the J6 ORF (7; E. V. Jones, C. Puckett, and B. Moss, manuscript in preparation). Some of the other minor bands may be premature termination products of the very long RNA polymerase message. The mRNA encoding the 30-kDa protein was selected by hybridization to M13 H42 and M13 H52 single-stranded DNA (Fig. 2A). Although both H5L and H6R are overlapped by M13 H41, only the latter is in the correct DNA strand (Fig. 2C). The predicted size of the H6R product, however, was only 22.3 kDa. Investigations into this discrepancy are described below.

When late RNAs that hybridized to the *Hin*dIII H fragment were translated, the most prominent polypeptides had apparent molecular weights of about 40,000, 35,000, and 19,000 (Fig. 2B). The latter appeared as a smear and might correspond to two distinct bands previously observed by Bajszar et al. (1). Because of their long and heterogeneous 3′ ends, late mRNAs may be selected by DNA fragments that lie considerably downstream of the coding region. In addition, networks formed by annealing of complementary late RNAs (6, 9) may result in selection by coding and noncoding DNA strands. Both of these problems are evident to varying degrees in Fig. 2B. Nevertheless, several polypeptides were unambiguously assigned to ORFs.

The 19-kDa polypeptide was synthesized by mRNA that hybridized specifically to M13 H11 DNA which overlaps ORFs J6R, H1L, and H2R (Fig. 2B and C). However, only H1L is in the proper DNA strand. Furthermore, the H1L ORF predicts a closely matching polypeptide of 19.7 kDa.

The 35-kDa polypeptide was made in greatest amount

TABLE 2. Amino acid (AA) composition of putative proteins derived from ORFs

Amino acid[a]	mol (%)						
	H1 (171 AA)	H2 (189 AA)	H3 (324 AA)	H4 (178 AA)	H5 (575 AA)	H6 (203 AA)	H7 (314 AA)
Ala	9 (5.2)	16 (8.4)	17 (5.2)	6 (3.4)	21 (3.6)	17 (8.4)	9 (2.9)
Arg	7 (4.1)	8 (4.2)	14 (4.3)	7 (3.9)	23 (4.0)	8 (3.9)	15 (4.8)
Asn	13 (7.6)	11 (5.8)	22 (6.8)	13 (7.3)	42 (7.3)	7 (3.4)	19 (6.0)
Asp	12 (7.0)	11 (5.8)	22 (6.8)	9 (5.0)	38 (6.6)	17 (8.4)	18 (5.7)
Cys	2 (1.2)	5 (2.6)	2 (0.6)	3 (1.7)	6 (1.0)	1 (0.5)	2 (0.6)
Gln	2 (1.2)	6 (3.2)	4 (1.2)	3 (1.7)	10 (1.7)	4 (2.0)	7 (2.2)
Glu	5 (2.9)	8 (4.2)	21 (6.5)	12 (6.7)	34 (5.9)	18 (8.9)	17 (5.4)
Gly	5 (2.9)	9 (4.7)	14 (4.3)	1 (0.6)	13 (2.3)	5 (2.5)	11 (3.5)
His	4 (2.3)	2 (1.1)	9 (2.8)	6 (3.4)	9 (1.6)	2 (1.0)	8 (2.5)
Ile	10 (5.8)	11 (5.8)	32 (9.8)	18 (10.1)	67 (11.6)	13 (6.4)	20 (6.3)
Leu	17 (9.9)	17 (8.9)	21 (6.5)	18 (10.1)	52 (9.0)	8 (3.9)	27 (8.6)
Lys	16 (9.3)	16 (8.4)	23 (7.1)	14 (7.8)	47 (0.2)	26 (12.8)	34 (10.8)
Met	9 (5.2)	7 (3.7)	11 (3.4)	3 (1.7)	10 (1.7)	4 (2.0)	9 (2.9)
Phe	6 (3.5)	11 (5.8)	24 (7.4)	10 (5.6)	38 (6.6)	3 (1.5)	21 (6.7)
Pro	7 (4.1)	3 (1.6)	11 (3.4)	9 (5.0)	13 (2.3)	8 (3.9)	10 (3.2)
Ser	13 (7.6)	13 (6.8)	11 (3.4)	10 (5.6)	48 (8.3)	20 (9.9)	21 (6.7)
Thr	10 (5.8)	12 (6.3)	22 (6.8)	9 (5.0)	26 (4.5)	20 (9.9)	21 (6.7)
Trp	0 (0.0)	4 (2.1)	4 (1.2)	1 (0.6)	3 (0.5)	1 (0.5)	2 (0.6)
Tyr	12 (7.0)	7 (3.7)	11 (3.4)	14 (7.8)	40 (6.9)	1 (0.5)	14 (4.4)
Val	12 (7.0)	12 (6.3)	29 (8.9)	12 (6.7)	35 (6.1)	20 (9.9)	29 (9.2)
Acidic	17 (9.9)	19 (10.0)	43 (13.2)	21 (11.7)	72 (12.5)	35 (17.2)	35 (11.1)
Basic	23 (13.4)	24 (12.6)	37 (11.4)	21 (11.7)	70 (12.2)	34 (16.7)	49 (15.6)
Aromatic	18 (10.5)	22 (11.6)	39 (12.0)	25 (14.0)	81 (14.1)	5 (2.5)	37 (11.7)
Hydrophobic	66 (38.4)	69 (36.3)	132 (40.6)	76 (42.5)	245 (42.5)	50 (24.6)	122 (38.7)
Mol wt	19,700	21,500	37,500	21,200	67,600	22,300	36,700

[a] Acidic, Asp plus Glu; basic, Arg plus Lys; aromatic, Phe plus Trp plus Tyr; hydrophobic, aromatic plus Ile plus Leu plus Met plus Val.

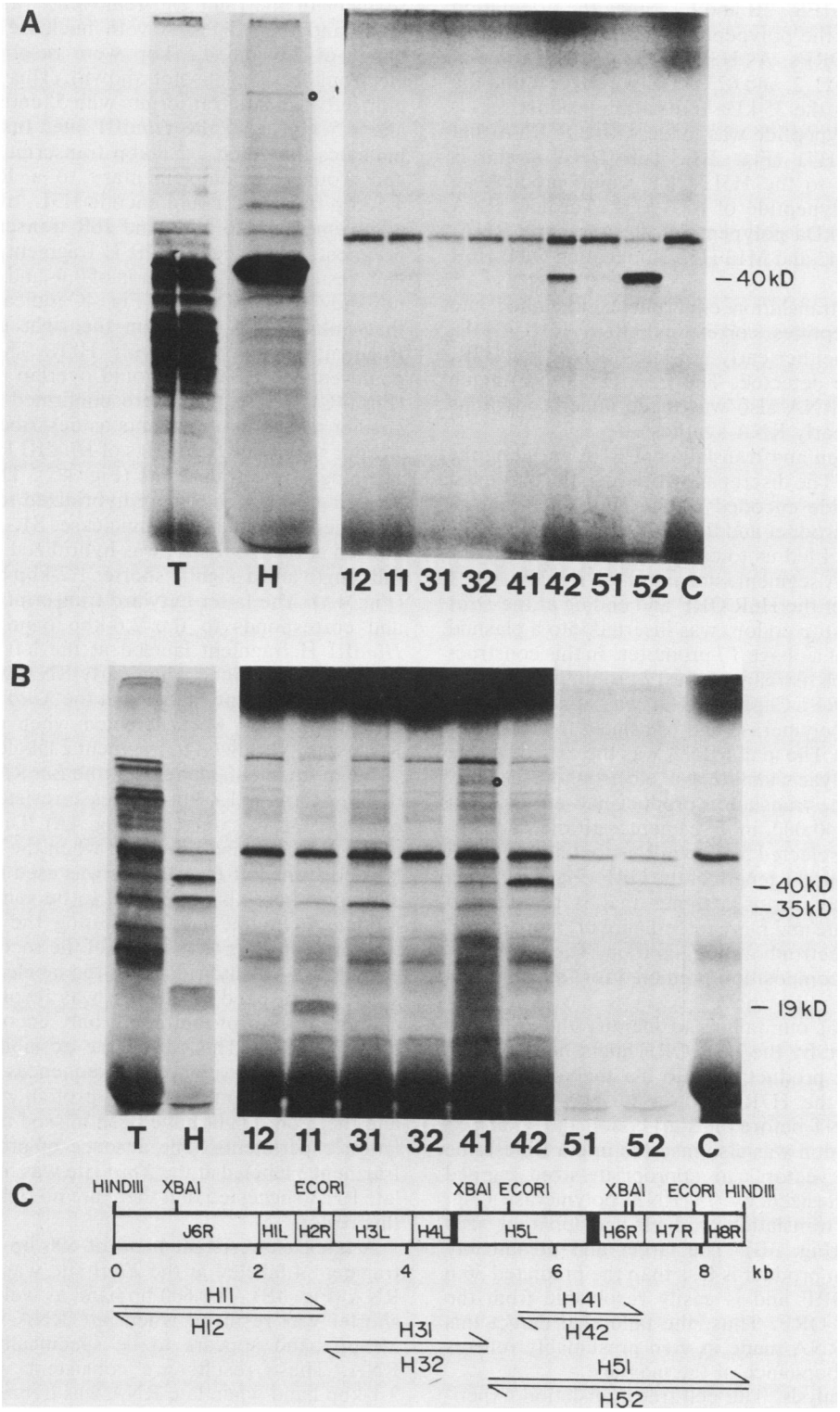

FIG. 2. Analysis of cell-free translation products of early and late mRNAs selected by hybridization to single-stranded subcloned fragments of *Hin*dIII H. (A) Translation products of early mRNA. (B) Translation products of late mRNA. (C) Diagram indicating ORFs and restriction fragments subcloned in M13mp18 or M13mp19 phage. Lanes: T, total unselected RNA; C, control without RNA; H, RNA selected by hybridization to both strands of entire *Hin*dIII H fragment. Numbers refer to RNA that hybridized to corresponding single-strand phage. The dots in (A) and (B) point out the faint 147- and 69-kDa bands, respectively. Autoradiographs of polyacrylamide gels are shown.

when the cell-free extract was programmed with RNA that hybridized to M13 H31 DNA which overlaps parts of H2R, H3L, H4L, and H5L (Fig. 2B and C). Since the orientation of H2R is incorrect, the polypeptide must be a product of one of the leftward ORFs. As H3L, H4L, and H5L predict polypeptides of 37.5, 21.2, and 67.6 kDa, we suggest that the first of these encodes the 35-kDa translation product.

A faint 69-kDa polypeptide was made with late RNA that hybridized to M13 H41 (Fig. 2B). This DNA strand is complementary only to the H5L ORF which predicts an appropriate-sized polypeptide of 67.6 kDa. The late RNA that encoded the 40-kDa polypeptide, like the early RNA, hybridized to M13 H42 and M13 H52, implicating ORF H6R (Fig. 2B and C).

In summary, the translation experiments identified late mRNAs and polypeptides corresponding to H1L, H3L, H5L, and H6R but neither early nor late mRNAs for H2R, H4L, and H7R were detected. The H6R ORF was unique since a translatable RNA also was made under conditions that permitted only early RNA synthesis.

In vitro transcription and translation of RNA encoding the H6R and H7R ORFs. The discrepancy between the predicted size of the polypeptide encoded by the H6R ORF and its cell-free translation product and the apparent absence of the adjacent H7R product led us to carry out additional experiments. A 700-bp DNA segment, starting at the *Eco*RV site 21 NT before the start of the H6R ORF and ending at the *Dra*I site 54 NT after the stop codon, was inserted into a plasmid vector downstream of a phage T7 promoter. In this construction, the first ATG downstream of the T7 promoter is the one starting the H6R ORF. Capped RNA was synthesized in vitro with T7 RNA polymerase and translated in a reticulocyte cell-free system. The major RNA was the expected size as measured by polyacrylamide gel electrophoresis (Fig. 3A). Significantly, the translation product had an apparent molecular weight of 40,000, in agreement with the polypeptide made by hybrid-selected in vivo mRNA but nearly twice the size predicted from the length of the ORF (Fig. 3B). Thus the polypeptide either forms a dimer that is resistant to sodium dodecyl sulfate and mercaptoethanol or, more likely, has an anomalous electrophoretic migration. The somewhat unusual amino acid composition predicted for this polypeptide was commented upon above.

We considered that our failure to identify the 36.7-kDa polypeptide predicted by the H7R ORF might be due to its comigration with the product of H6R. To analyze the polypeptide product of the H7R ORF, a 1.1-kbp *Alu*I DNA segment starting 51 NT before the start codon and ending 53 NT after the stop codon was also inserted into a Bluescribe plasmid expression vector. An appropriate-sized capped transcript was synthesized by T7 RNA polymerase (Fig. 3A), and the in vitro translation products had apparent sizes of 32 and 25 kDa (Fig. 3B). The larger and presumably full-length translation product is less than the predicted 36.6 kDa for the H7R ORF and is easily resolvable from the product of the H6R ORF. Thus, the failure to detect this polypeptide with mRNA made in vivo presumably reflects an absence or low abundance of the message.

Transcriptional analysis. The cell-free translation experiments indicated that the majority of genes in the *Hin*dIII H fragment are expressed at late times. Previous studies have shown that the extreme length and 3′ heterogeneity of late transcripts prevent the detection of individual species by electrophoretic blotting procedures (11, 22, 34). We considered, however, that the length and overlapping of late transcripts might provide an advantage for preliminary mapping of the 5′ ends of several late RNAs simultaneously. When the left end-labeled *Hin*dIII H fragment, with a small segment of the right end removed, was hybridized to late RNA and then digested with nuclease S1, two protected bands of 2.6 and 4.5 kbp were resolved by agarose gel electrophoresis (data not shown). This suggested overlapping leftward late transcripts with 5′ ends approximately 2.6 and 4.5 kbp from the *Hin*dIII site. Inspection of Fig. 4C indicates that such a 2.6-kbp transcript could encode H1L and would be complementary to a J6R transcript. The 4.5-kbp transcript could encode H3L, overlap H1L, and be complementary to H2R and J6R transcripts. A similar experiment, using the *Hin*dIII H fragment labeled at the right end, resulted in the formation of 0.4- and 2.1-kbp bands (data not shown). Inspection of Fig. 4C indicates that a transcript that initiates 0.4 kbp from the right end of the *Hin*dIII fragment could encode only H8R. The 2.1-kb transcript could encode H6R and would overlap both H7R and H8R (Fig. 4C). These data were confirmed and extended with smaller restriction fragments as described below.

Four subcloned fragments of *Hin*dIII H were used for finer mapping of the transcripts (Fig. 4C). The DNA fragments, labeled at either end, were hybridized to early or late RNA and then digested with nuclease S1. When fragment 2, labeled at the *Xba*I site, was hybridized to late RNA, both a full-length and a slightly shorter 1.7-kbp band were protected (Fig. 4A). The latter leftward transcript could express H1L and corresponds to the 2.6-kbp band obtained with the *Hin*dIII H fragment labeled at the left end. No protected bands were obtained when early RNA was used for hybridization to fragment 2 labeled at the *Xba*I site. In addition, no protected bands were detected when either late or early RNA was hybridized to fragment 2 labeled at the *Eco*RI site.

When fragment 3 labeled at the *Eco*RI site was hybridized to late RNA, a 1.4-kbp nuclease-resistant band that could express H3L was detected (Fig. 4A). This band corresponds to the 4.5-kbp band produced when the *Hin*dIII H fragment labeled at the left *Hin*dIII site was used for hybridization. In addition, there also appeared to be some protection of the full-length fragment 3, suggesting the presence of a late transcript starting to the right of the second *Xba*I site. When early RNA was hybridized to the labeled fragment, a nuclease-resistant band of 0.6 kbp was resolved (Fig. 4A), indicating a transcript that could only encode the distal half of the H3L ORF. This raised the possibility of two proteins with overlapping amino acid sequences expressed from the same ORF. The putative early protein would be only 13 kDa and therefore might have been missed in the cell-free translation experiments. The absence of protected bands when fragment 3 labeled at the *Xba*I site was hybridized to early or late RNA suggested that there are no rightward transcripts in this region.

A nuclease-resistant band of 600 bp was obtained when fragment 4 labeled at the *Xba*I site was hybridized to early RNA (Fig. 4B). The 600-bp band as well as one about 50 bp shorter was resolved when late RNA was used. Thus the 550-bp band appears to be specifically formed with late RNA. These results are consistent with detection of a 2.1-kbp band when late RNA was hybridized to the *Hin*dIII H fragment labeled at the right *Hin*dIII site. With fragment 4 labeled at the *Eco*RI site, late RNA protected a 1-kbp fragment (Fig. 4B) which could encode ORF H5.

Late RNA, when hybridized to fragment 5 labeled at the *Hin*dIII site, protected a 0.4-kbp fragment which could encode ORF H8R (Fig. 4B). No protected bands were detected when early RNA was used.

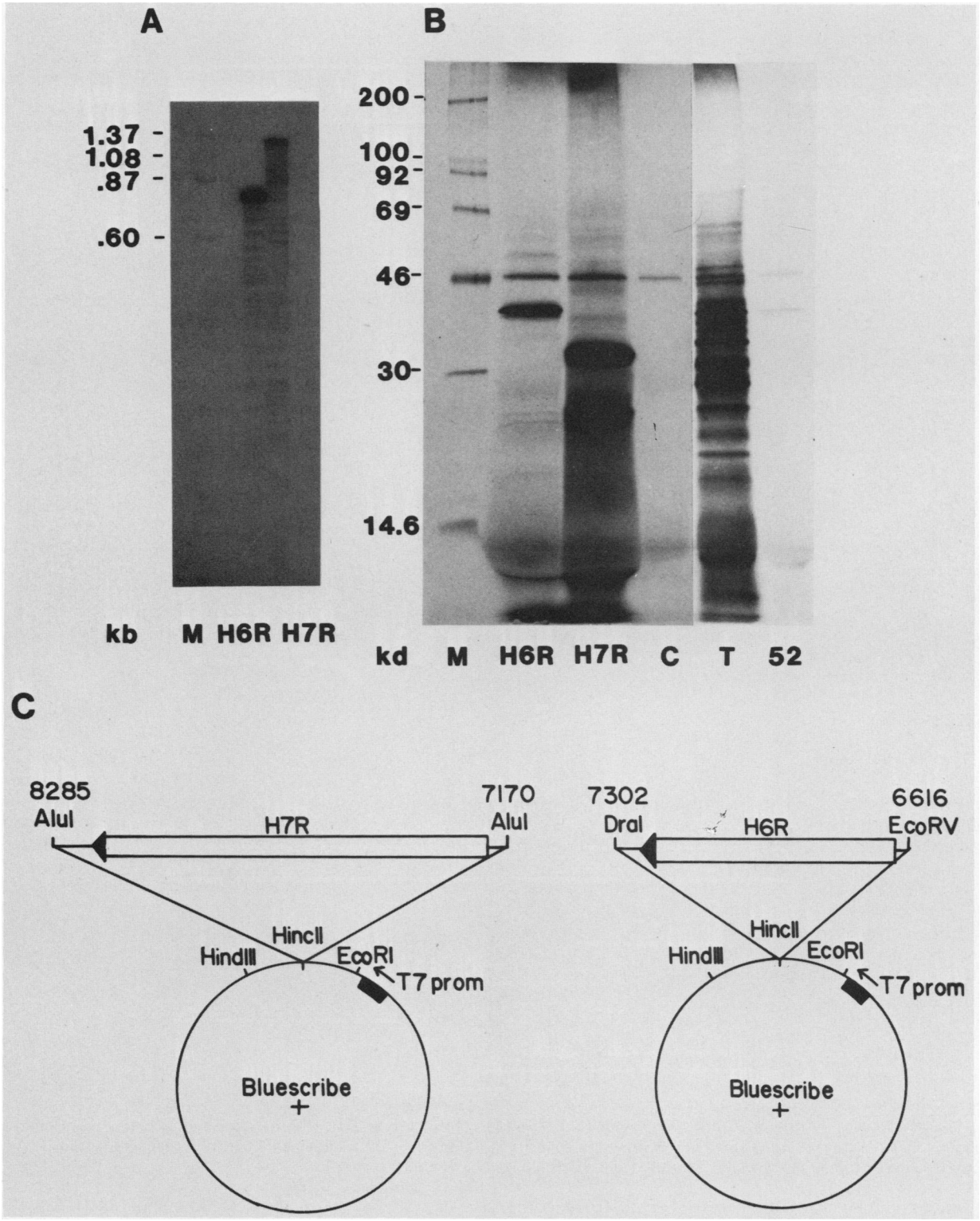

FIG. 3. In vitro transcription and translation of cloned DNA segments. (A) Polyacrylamide gel electrophoresis of ^{32}P-labeled RNA synthesized in vitro by T7 RNA polymerase. (B) Polyacrylamide gel electrophoresis of cell-free translation products of in vitro RNA. (C) Recombinant plasmids containing H7R or H6R ORFs used as a template for in vitro RNA synthesis. The nucleotides defining the DNA fragments inserted into the Bluescribe vector are indicated. Lanes: M, marker DNA or proteins; T, total RNA; C, control without RNA; H6R or H7R, recombinant plasmid containing corresponding ORF used as template; 52, M13 phage indicated in Fig. 2 used for selection of RNA.

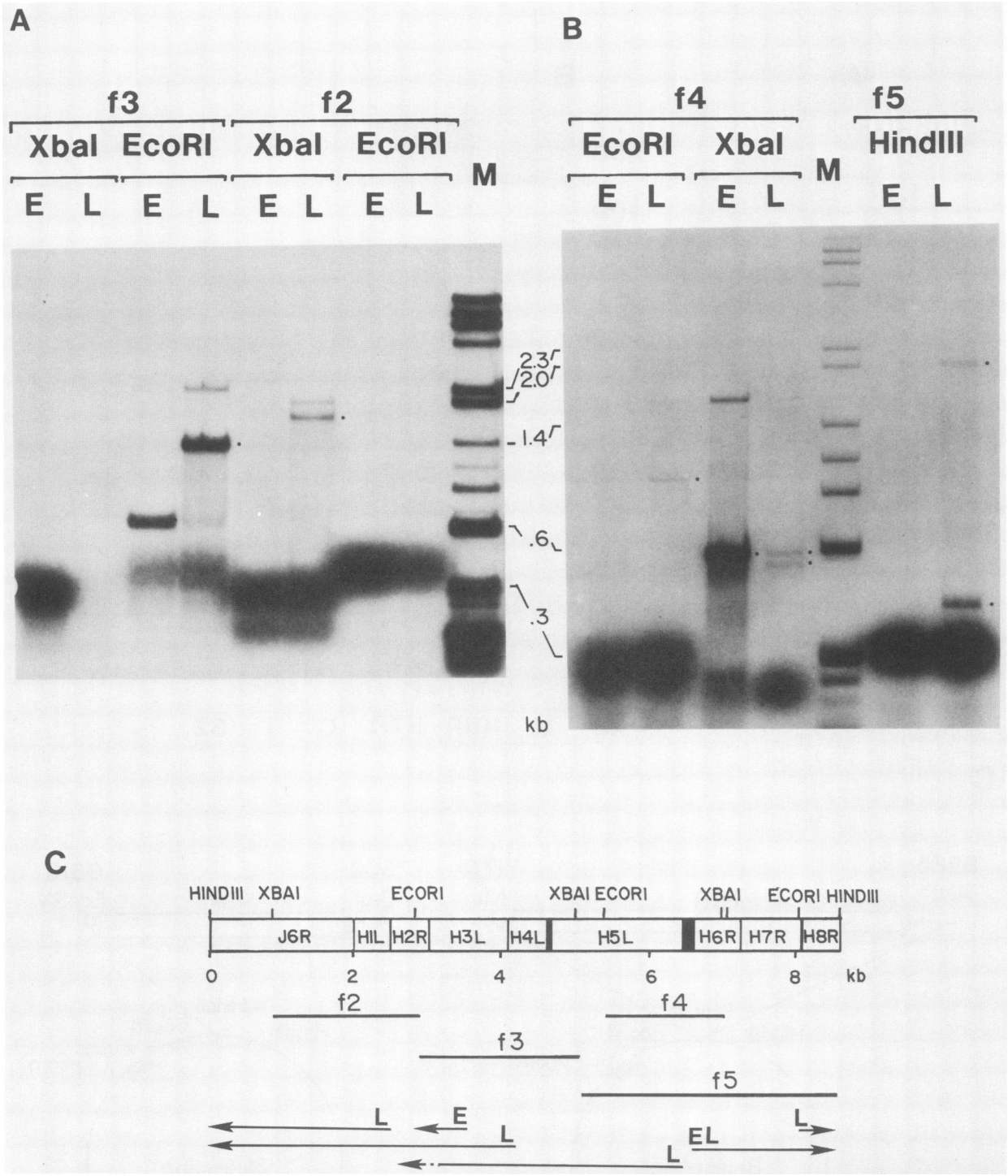

FIG. 4. Agarose gel electrophoresis of nuclease S1-protected DNA-RNA hybrids. (A and B) Autoradiographs of gels following electrophoresis of nuclease S1-resistant RNA-DNA hybrids. E and L refer to early or late RNA used for hybridization. Restriction sites of asymmetrically labeled DNA fragments are indicated. (B) Diagram showing ORFs, DNA fragments (f2 to f5) used as hybridization probes, and late (L) and early (E) RNA species (arrows) deduced from nuclease S1 analysis.

In summary, nuclease S1 analysis served to map the approximate positions of five late transcripts with 5′ ends that closely precede ORFs H1L, H3L, H5L, H6R, and H8R. One early transcript has a 5′ end, suggesting that it also could express H6R. The finding of early and late transcriptional start sites for the H6R ORF was consistent with the synthesis of the 40-kDa translation product with early and late RNA. Unexpectedly, we detected an early transcript with a 5′ end that mapped to the middle of H3L and which could encode the distal half of that ORF. We did not find 5′ ends of transcripts that corresponded to H2R, H4R, or H7R, just as we could not find translation products for them.

Analysis of the H6R transcript by agarose gel electrophoresis. The presence of an early H6R transcript was confirmed by gel electrophoresis of RNA synthesized in the presence of

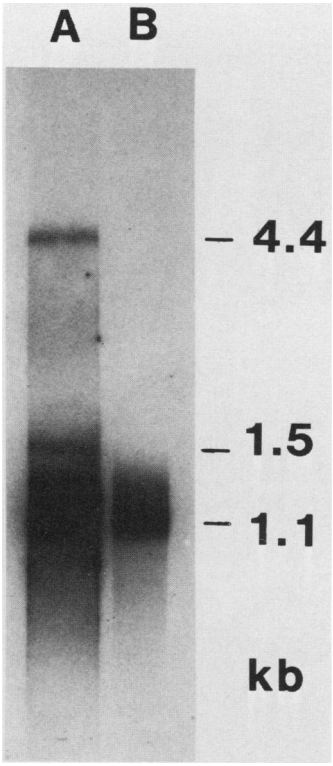

FIG. 5. Blot hybridization analysis of early RNAs. Cytoplasmic RNA from cells infected with vaccinia virus in the presence of cycloheximide was subjected to electrophoresis on a 1% agarose gel and then transferred to a nitrocellulose membrane. ^{32}P-labeled *Hin*dIII H fragment (A) or f4 described in the legend to Fig. 3 (B) was hybridized to blotted RNAs.

cycloheximide. The RNA was transferred to a nitrocellulose membrane and probed with ^{32}P-labeled M13 H41 replicative-form DNA. A prominent broad band of 1.1 kilobases (kb), appropriate in size for the mRNA encoding the H6R ORF, was detected. When the entire *Hin*dIII H DNA was used as a probe, two additional early transcripts were detected (Fig. 5). The 4.4-kbp band corresponds to the long mRNA encoding J6R, the large subunit of RNA polymerase (7). The minor 1.5-kbp band could correspond to the mRNA detected by nuclease S1 analysis starting within the H3L ORF.

Fine mapping of the 5′ ends of early and late transcripts. The proximity of the 5′ ends of transcripts to the putative translation start codons of the ORFs was determined by high-resolution nuclease S1 analysis, using small DNA hybridization probes, and electrophoresis of the nuclease-protected fragments beside Maxam-Gilbert sequence ladders (Fig. 6). In most cases mung bean nuclease was used to avoid or minimize nibbling artifacts (15). The 5′ ends, determined from the positions of the major bands, are indicated next to the genome sequence in Fig. 1. For the ORFs designated H1, H3, and H8, the 5′ end of the RNA appeared to precede the ATG codon by only 1 or 2 NT. In the case of H5 the apparent RNA start site preceded the ATG by 7 or 8 NT. The early and late RNA start sites of H6R were separated from each other by about 35 NT.

DISCUSSION

The present study provides the longest vaccinia virus DNA sequence for which detailed transcriptional and translational mapping has been carried out. The *Hin*dIII H DNA fragment was chosen because it contains several late class genes, for which there is little information regarding organization and regulation, and because of its location in the central highly conserved region of the genome adjacent to a block of DNA on the left that had already been sequenced (7, 26, 33, 34). In Fig. 7, we have compiled present and previously published data for ORFs in the three contiguous *Hin*dIII fragments and, where the information is available, have indicated transcription and translation products. Two ORFs, J4R and J6R, specify RNA polymerase subunits (7, 24; Jones et al., in preparation), J2R is the thymidine kinase gene (17, 33), and L4R encodes a major proteolytically processed core polypeptide (34, 35). The functions of the other putative genes are unknown. Although H3L and H7R have short regions of sequence similarity to the precursor spike glycoprotein of vesicular stomatitis virus and to ribonucleotide reductase, respectively, there is no additional information to support functional or evolutionary relationships.

In this 16-kbp segment of the vaccinia virus genome, portions of both DNA strands are transcribed, although rightward products are predominant. The direction of ORFs do not correlate with early or late expression. Some clustering of early genes in the *Hin*dIII J fragment and late genes in the *Hin*dIII H fragment is apparent. The ORFs in the *Hin*dIII H fragment are closely spaced. Examples of ORFs with overlapping stop codons (J6R and HL1) and exactly contiguous stop and start codons (H4L and H3L, H7R and H8R) were observed. There are examples of more extensive ORF overlaps in the *Hin*dIII J and L fragments (26).

For some ORFs, e.g., H2, H4, and H7, we could not find evidence of specific transcripts. Nevertheless, it would be inappropriate at this stage to assume that these genes are not expressed since the transcripts may be present at low levels under the conditions and times used for isolation of early and late RNA. It is also possible that these ORFs are expressed exclusively on polycistronic messages, although we know of no precedent for such a phenomenon in a eucaryotic virus system.

The most interesting correlation made in this report concerns the sequence around the ATG start codons of late genes and their close proximity to RNA start sites. The first sequenced vaccinia virus late genes, encoding a 28-kDa core polypeptide precursor (34) and an 11-kDa core polypeptide (3), both contained TAAATG sequences at the start of the ORFs. A related sequence, TAAATAATG, was found at the start of a third late ORF encoding still another core polypeptide precursor (28). In each case, the RNA start sites were mapped within several nucleotides of the ATG. In the present report, we demonstrate that four additional late ORFs also have the sequence TAAATG and that the 5′ ends of transcripts map within or just upstream of the run of three A's (Fig. 8). By contrast, this sequence has not been found at the start of ORFs listed in Fig. 8 that are expressed early in infection (7, 13, 32, 33). The most similar sequences at the start of early ORFs are GAAATG for DNA polymerase (13) and TAAAATG for the large subunit of RNA polymerase (7). Two vaccinia virus genes have been identified that are preceded by both early and late RNA start sites: the late RNA start site of the 7.5-kDa gene (31) has two repeating TAAATA sequences and the one for the H6R ORF contains the sequence TAAATT. Several additional examples of TAAATG sequences are present in the *Hin*dIII L fragment (26); however, these genes have not been classified and the RNA start sites were not mapped.

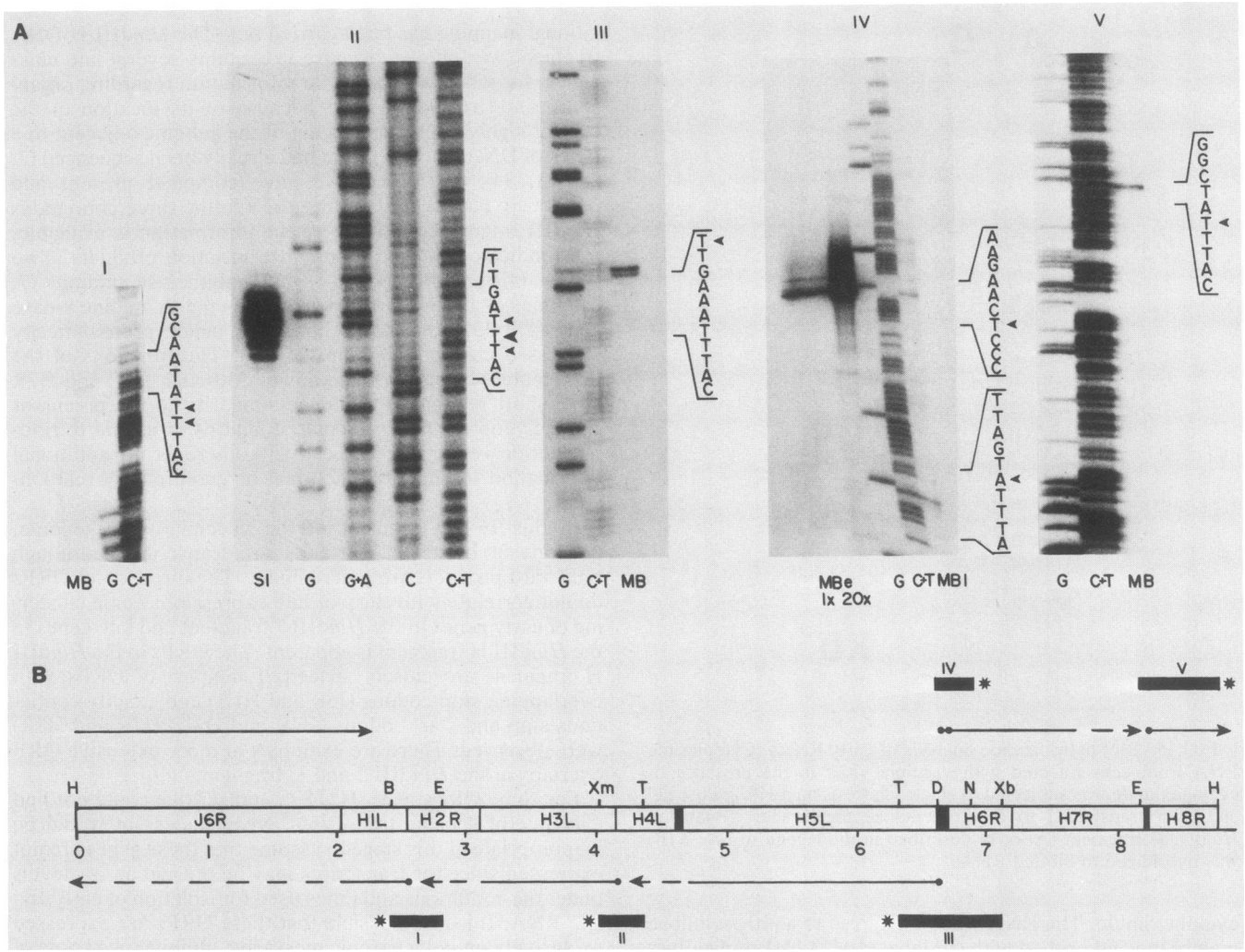

FIG. 6. High-resolution analysis of 5′ ends of transcripts. (A) Late RNA was hybridized to asymmetrically labeled DNA fragments indicated by Roman numerals and digested with either mung bean nuclease (MB) or nuclease S1 (S1). The nuclease-resistant DNA was analyzed by electrophoresis alongside the indicated (G, C+T, G+A, C) Maxam-Gilbert sequence leaders prepared with the same DNA used for hybridization. The DNA sequences surrounding the endonuclease-resistant bands are shown, and the positions of the major bands are indicated by asterisks in Fig. 1. (B) ORFs are indicated above a scale in kilobase pairs. Restriction endonuclease sites are abbreviated: H, *Hin*dIII; B, *Bam*HI; E, *Eco*RI; Xm, *Xmn*I; T, *Taq*I; D, *Dra*I; N, *Nci*I; Xb, *Xba*I. Arrows indicate transcripts and filled blocks are restriction fragments used for hybridization. The asterisk indicates the asymmetrically labeled end.

The TAAATG sequence and the short RNA leader are consistent features of late genes and may in part distinguish them from early genes. A-T-rich segments centered about 15 bp upstream of the ATG also is a common feature; however, early genes may have even more highly A-T-rich sequences upstream of their RNA start sites. At this time, we are uncertain whether the TAAATG sequence has a transcriptional or translational function but would not be surprised if it had elements for both. The TAAAT may have a transcriptional role since it is also present at the late start site of the early/late promoters which lack the G to form the translation initiation codon.

The sequences surrounding the translation initiation codons of eucaryotic mRNAs have been compiled by Kozak (18). The most conserved nucleotides are the purine at −3 and the G at +4, relative to the A of the ATG. Only 2 of 211 eucaryotic mRNAs have a T in the −3 position. By contrast, six of seven late vaccinia virus ORFs have a T (Fig. 8). Mutagenesis studies indicated that a T in the −3 position lowers the efficiency of translation in vivo; however, this effect is partially mitigated by a G in the −4 position (19). It seems significant, therefore, that the majority of vaccinia virus late transcripts have a G in this position (Fig. 8) and are translatable in reticulocyte lysates. Whether the expression of late mRNAs is enhanced because of alterations in the translation system following vaccinia virus infection remains to be determined. In some cases, the TAA stop codon immediately precedes the ATG initiation codon, a proximity that could merely represent an extreme compactness of DNA sequences. In other cases, however, additional stop codons precede the TAAATG, making this explanation insufficient. Recent studies indicate that reinitiation of translation can occur following a stop codon in eucaryotic cells. Kozak (19) found that a purine at −3 and a G at +4 also enhance reinitiation and that the distance between the stop codon and ATG is not critical. Nevertheless, the TAAATG sequence was not specifically tested and it remains possible that this sequence is a favorable context for reinitiation to

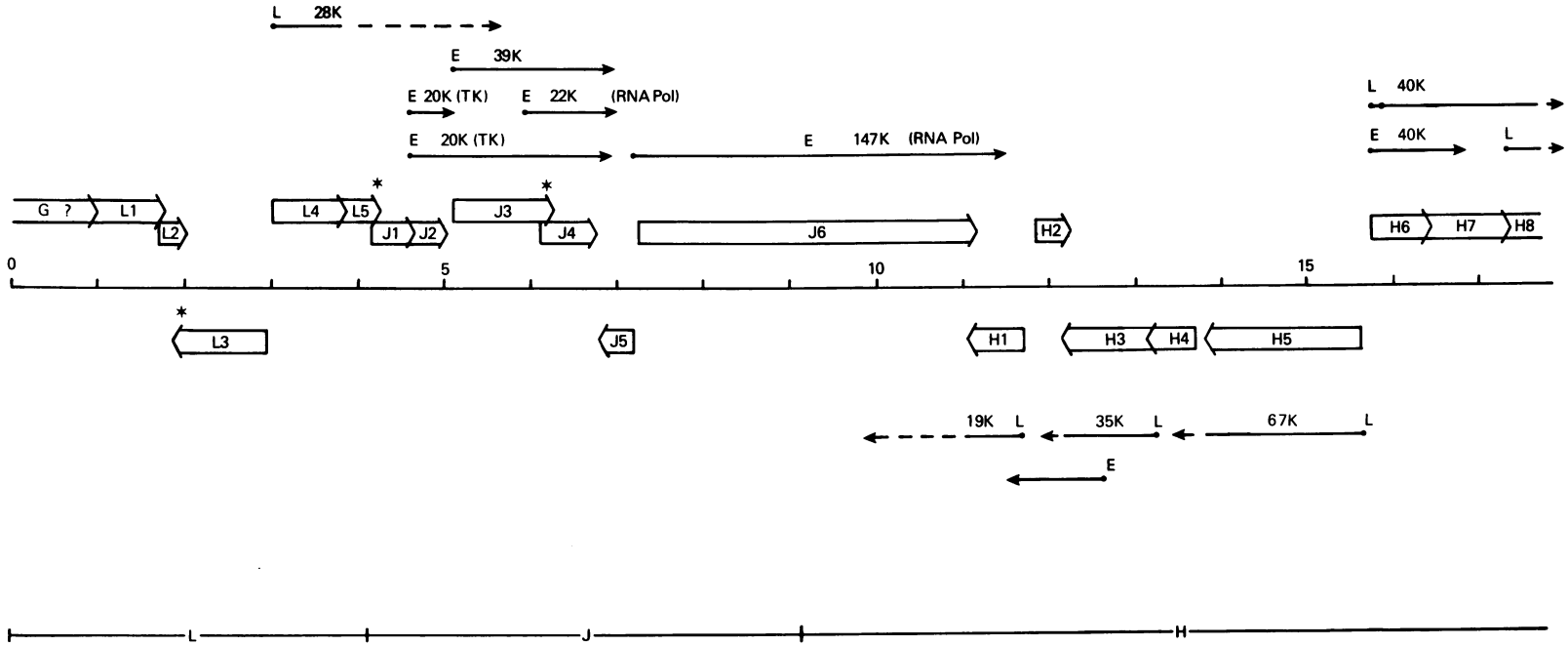

FIG. 7. Map of *Hin*dIII L, J, and H fragments of the vaccinia virus genome. The data used for this map were compiled from the present and previous (1, 25, 33) reports. Blocks, ORFs; arrows, transcripts; filled circles, 5′ ends of transcripts; E, early; L, late. The estimated molecular weights of translation products mapped to regions of the genome are indicated above transcripts.

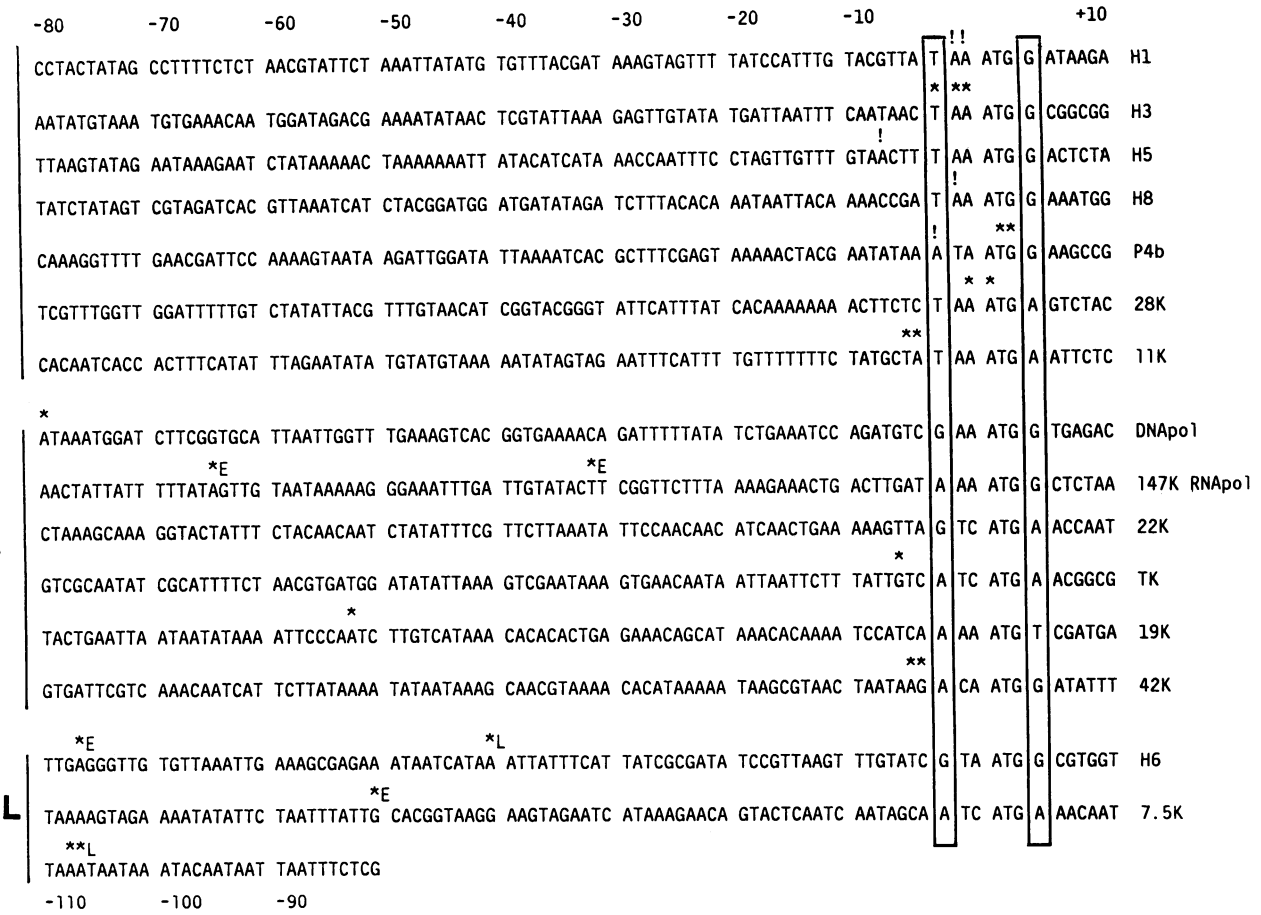

FIG. 8. Nucleotide sequences around translational and transcriptional start sites of vaccinia virus genes. The +1 position is the A of the ATG starting the ORF. Nucleotides in the −3 and +4 positions are boxed. Asterisks and exclamation marks indicate the major 5′ ends of transcripts determined by DNA-RNA hybridization and digestion with nuclease S1 or mung bean nuclease, respectively. Sequences are grouped into those expressed early (E), late (L), and early and late (E/L). Symbols on the extreme right indicate ORF or other designations for the vaccinia virus gene. Data are from present report and references 3, 7, 13, 28, and 31 to 34.

take place. In view of the characteristic long readthrough transcripts made at late times after vaccinia virus infection, reinitiation could be quite significant. Reinitiation could account for the expression at late times of low levels of promoterless genes introduced into the vaccinia virus genome (21).

The second unusual feature of late transcripts is their extremely short leader sequence. Mung bean and nuclease S1 mapping placed the 5′ end at the first or second A (−2 or −1 position relative to the ATG) within the TAAATG sequence in several cases and just a few nucleotides upstream in others. Obviously, if the 5′ end of the mRNA is in the −1 or −2 position, the translational role of the −3 nucleotide is moot. Although the nuclease experiments were done under relatively low-temperature and high-salt conditions to minimize nibbling, we are not entirely confident of such indirect methods of 5′-end analysis. Repeated attempts to use primer extension mapping procedures, however, were unsatisfactory. Our failure cannot be attributed solely to technique, since we had no trouble mapping the ends of early mRNAs by primer extension. We suspect that the problem is caused by the large amount of complementary late RNA that competes with or strand displaces shorter DNA primers. The start of many late RNAs 1 or 2 NT upstream of the ATG, however, is supported by direct analysis of the methylated ends of bulk late RNA (5). The ends were shown to be predominantly $m^7Gpppm^6A^mpA^m$, whereas early mRNAs had more G than A in the first nucleotide following the m^7G and all four NT were present in the following position. The presence of consecutive A residues as the first two transcriptionally added nucleotides of most late RNAs is consistent with initiation at the −2 or −1 position of the conserved TAAATG sequence.

Late mRNAs have two other unusual features: the presence of long heterogeneous 3′ ends (11, 22, 34) and the existence of large amounts of RNAs capable of forming extensive double-stranded networks (6, 9). The latter presumably results from transcriptional readthrough of oppositely oriented late genes. It might be anticipated that such antisense RNAs would result in hybridization arrest of translation. Perhaps the very short leader sequences of late mRNAs is an adaptation to minimize this inhibitory effect.

ACKNOWLEDGMENTS

We thank S. Broyles for providing sequence data for the left end of the *Hin*dIII H fragment, A. Davison for help in displaying DNA sequences, and N. Cooper for technical assistance.

LITERATURE CITED

1. **Bajszar, G., R. Wittek, J. P. Weir, and B. Moss.** 1983. Vaccinia virus thymidine kinase and neighboring genes: mRNAs and polypeptides of wild-type virus and putative nonsense mutants. J. Virol. **45**:62–72.
2. **Belle Isle, H., S. Venkatesan, and B. Moss.** 1981. Cell-free translation of early and late mRNAs selected by hybridization to cloned DNA fragments derived from the left 14 million to 72 million daltons of the vaccinia virus genome. Virology **112**:306–317.
3. **Bertholet, C., R. Drillien, and R. Wittek.** 1985. One hundred base pairs of 5′ flanking sequence of a vaccinia virus late gene are sufficient to temporally regulate late transcription. Proc. Natl. Acad. Sci. USA **82**:2096–2100.
4. **Birnboim, H. C., and J. Doly.** 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res. **7**:1513–1523.
5. **Boone, R. F., and B. Moss.** 1977. Methylated 5′-terminal sequences of vaccinia virus mRNA species made at early and late times after infection. Virology **79**:67–80.
6. **Boone, R. F., R. P. Parr, and B. Moss.** 1979. Intermolecular duplexes formed from polyadenylylated vaccinia virus RNA. J. Virol. **30**:365–374.
7. **Broyles, S., and B. Moss.** 1986. Homology between RNA polymerases of poxviruses, prokaryotes, and eukaryotes: nucleotide sequence and transcriptional analysis of vaccinia virus genes encoding 147-kDa and 22-kDa subunits. Proc. Natl. Acad. Sci. USA **83**:3141–3145.
8. **Cochran, M. A., C. Puckett, and B. Moss.** 1985. In vitro mutagenesis of the promoter region for a vaccinia virus gene: evidence for tandem early and late regulatory signals. J. Virol. **53**:30–37.
9. **Colby, C., C. Jurale, and J. R. Kates.** 1971. Mechanism of synthesis of vaccinia virus double-stranded ribonucleic acid in vivo and in vitro. J. Virol. **7**:71–76.
10. **Cooper, J. A., and B. Moss.** 1979. *In vitro* translation of immediate early, early and late classes of RNA from vaccinia virus-infected cells. Virology **96**:368–380.
11. **Cooper, J. A., R. Wittek, and B. Moss.** 1981. Extension of the transcriptional and translational map of the left end of the vaccinia virus genome to 21 kilobase pairs. J. Virol. **39**:733–745.
12. **DeFilippes, F. M.** 1982. Restriction enzyme mapping of vaccinia virus DNA. J. Virol. **43**:136–149.
13. **Earl, P., E. V. Jones, and B. Moss.** 1986. Homology between DNA polymerases of poxviruses, herpesviruses, and adenoviruses: nucleotide sequence of the vaccinia virus DNA polymerase gene. Proc. Natl. Acad. Sci. USA **83**:3659–3663.
14. **Golini, F., and J. R. Kates.** 1985. A soluble transcription system derived from purified vaccinia virions. J. Virol. **53**:205–213.
15. **Green, M. R., and R. G. Roeder.** 1980. Definition of a novel promoter for the major adenovirus-associated viral mRNA. Cell **22**:231–242.
16. **Henikoff, S.** 1984. Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene **28**:351–359.
17. **Hruby, D. E., R. A. Maki, D. B. Miller, and L. A. Ball.** 1983. Fine structure analysis and nucleotide sequences of the vaccinia virus thymidine kinase gene. Proc. Natl. Acad. Sci. USA **80**:3411–3415.
18. **Kozak, M.** 1984. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucleic Acids Res. **12**:857–872.
19. **Kozak, M.** 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukarotic ribosomes. Cell **44**:283–292.
20. **Lipman, D. J., and W. R. Pearson.** 1985. Rapid and sensitive protein similarity searches. Science **227**:1435–1441.
21. **Mackett, M., G. L. Smith, and B. Moss.** 1984. General method for the production and selection of infectious vaccinia virus recombinants expressing foreign genes. J. Virol. **49**:857–864.
22. **Mahr, A., and B. E. Roberts.** 1984. Arrangement of late RNAs transcribed from a 7.1-kilobase *Eco*RI vaccinia virus DNA fragment. J. Virol. **49**:510–520.
23. **Messing, J., and J. Viera.** 1982. A new pair of M13 vectors for selecting either strand of double digest restriction fragments. Gene **19**:269–276.
24. **Morrison, D. K., J. K. Carter, and R. W. Moyer.** 1985. Isolation and characterization of monoclonal antibodies directed against two subunits of rabbit poxvirus-associated DNA-directed RNA polymerase. J. Virol. **55**:670–680.
25. **Moss, B.** 1985. Replication of poxviruses, p. 685–703. *In* B. N. Fields, D. M. Knipe, and R. M. Chanock (ed.), Virology. Raven Press, New York.
26. **Pluciennicak, A., E. Schroeder, G. Zettlemeissl, and R. E. Streeck.** 1985. Nucleotide sequence of a cluster of early and late genes in a conserved segment of the vaccinia virus genome. Nucleic Acids Res. **13**:985–998.
27. **Rohrmann, G., and B. Moss.** 1985. Transcription of vaccinia virus early genes by a template-dependent soluble extract of purified virions. J. Virol. **56**:349–355.
28. **Rosel, J., and B. Moss.** 1985. Transcriptional and translational mapping and nucleotide sequence analysis of a vaccinia virus gene encoding the precursor of the major core polypeptide 4b. J. Virol. **56**:830–838.
29. **Sanger, F., A. R. Coulson, B. J. Barrell, A. J. H. Smith, and B. A. Roe.** 1980. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. J. Mol. Biol. **143**:161–178.
30. **Slabaugh, M. B., T. L. Johnson, and C. K. Mathews.** 1984. Vaccinia virus induces ribonucleotide reductase in primate cells. J. Virol. **52**:507–514.
31. **Venkatesan, S., B. M. Baroudy, and B. Moss.** 1981. Distinctive nucleotide sequences adjacent to multiple initiation and termination sites of an early vaccinia virus gene. Cell **25**:805–813.
32. **Venkatesan, S., A. Gershowitz, and B. Moss.** 1982. Complete nucleotide sequence of two adjacent early vaccinia virus genes located within the inverted terminal repetition. J. Virol. **44**:637–646.
33. **Weir, J. P., and B. Moss.** 1983. Nucleotide sequence of the vaccinia virus thymidine kinase gene and the nature of spontaneous frameshift mutations. J. Virol. **46**:530–537.
34. **Weir, J. P., and B. Moss.** 1984. Regulation of expression and nucleotide sequence of a late vaccinia virus gene. J. Virol. **51**:662–669.
35. **Weir, J. P., and B. Moss.** 1985. Use of a bacterial expression vector to identify the gene encoding a major core protein of vaccinia virus. J. Virol. **56**:534–540.
36. **Winberg, G., and M.-L. Hammarskjold.** 1980. Isolation of DNA from agarose gels using DEAE paper. Application to restriction site mapping of adenovirus type 16 DNA. Nucleic Acids Res. **8**:253–264.
37. **Wittek, R., A. Menna, D. Schumperli, S. Stoffel, H. Muller, and R. Wyler.** 1977. *Hin*dIII and *Sst*I restriction sites mapped on rabbit poxvirus and vaccinia virus DNA. J. Virol. **23**:669–678.