Nucleotide sequence and genetic organization of a 7.3 kb region (map unit 47 to 52.5) of Autographa californica nuclear polyhedrosis virus fragment EcoRI-C

Marcel Kool, René Broer, Douwe Zuidema, Rob W. Goldbach and Just M. Vlak*

Department of Virology, Wageningen Agricultural University, Binnenhaven 11, 6709 PD Wageningen, The Netherlands

The nucleotide sequence and genetic organization of an 7297 bp region within the EcoRI-C fragment of Autographa californica multiple nucleocapsid nuclear polyhedrosis virus (AcMNPV) are presented. Eight putative open reading frames were found and their respective amino acid sequences compared with a number of data libraries. ORF 1227 corresponded with gp41 and its predicted protein sequence was found to be 55 amino acids longer at its C terminus than reported previously. Moreover the main part of the ORF 1227 product, including the additional 55 amino acids, showed a high degree of homology with protein p40 of Helicoverpa zea single nucleocapsid nuclear polyhedrosis virus (HzSNPV). Three other ORFs in the analysed AcMNPV region showed homology with ORFs in the HzSNPV sequence, indicating that the general organization of this region is similar in both viruses. However one ORF found in the AcMNPV sequence was absent from the corresponding HzSNPV sequence.

Introduction

The genus nuclear polyhedrosis virus (NPV) of the family Baculoviridae includes the single (S) and multiple (M) nucleocapsid NPVs (Francki et al., 1991). The MNPV of Autographa californica (AcMNPV) represents the prototype baculovirus and its DNA is approximately 130 kbp in size. Baculoviruses contain circular, dsDNA genomes and their gene organization and sequence are being investigated (see Blissard & Rohrmann, 1990; Kool & Vlak, 1993 for review). About 60% of the AcMNPV genome has been sequenced. The replication of baculoviruses can be studied in much detail in the insect cell culture system.

Serial, undiluted passage of AcMNPV in Spodoptera frugiperda insect cell culture results in the generation of defective viruses (Kool et al., 1991; Wickham et al., 1991; Lee & Krell, 1992). These defective viruses are characterized by the presence of major genomic deletions and their dependence on helper virus for replication in insect cells. In addition, these defective viruses have interfering properties (Kool et al., 1991). DNA analysis showed that in these defective viruses up to 43% of the standard viral genome, a segment extending from map unit (m.u.) 1.7 to 45.0 on the physical map of AcMNPV, has been deleted. Further analysis of the defective genomes in serially passaged virus preparations revealed the presence of many different classes of deletion mutants all with their deletions ending in the EcoRI-C fragment (M. Kool, unpublished).

From the observation that some regions of the AcMNPV genome are maintained in defective viral DNA molecules it is logical to assume that they contain cis-acting elements necessary for DNA replication and/or encapsidation. Regions with highly repetitive (hr) sequences in the retained fragments could serve as an origin of DNA replication (Pearson et al., 1992; Kool et al., 1993a, b). These regions are also implicated as enhancers of transcription (Guarino et al., 1986). Encapsidation signals have not yet been described, but it was postulated that they are present near the junctions in the defective genomes (Kool et al., 1991). One junction is within the EcoRI-C fragment.

To investigate the genetic functions of the EcoRI-C fragment in the defective genomes and their possible role in the generation of these genomes, the sequence of this fragment needs to be determined. Whitford & Faulkner (1992b) sequenced the AcMNPV SstII-M fragment, which is located entirely within the EcoRI-C fragment and which contained an open reading frame (ORF) coding for a structural glycoprotein (gp41) of polyhedra-derived virions (Whitford & Faulkner, 1992a). Here, we present the nucleotide sequence of 7297 bp of the right...
Fig. 1. EcoRI restriction map of the AcMNPV genome and a partial restriction map of the region that has been sequenced. The position of the ORFs and the direction of transcription are indicated. E in hr3, EcoRI.

part of the EcoRI-C fragment, extending from the left SstI site of SstI-H to the first EcoRI site of the hr3 region of AcMNPV. Eight putative ORFs were found. Evidence is presented that the overall organization of this region is similar in both MNPVs and SNPVs.

Methods

Virus. The E2 strain of AcMNPV (Smith & Summers, 1978) was used.

Plasmid constructions. AcMNPV fragment EcoRI-C was cloned into plasmid pBR322. Overlapping subfragments, generated by a number of restriction enzymes, were cloned into pUC19 or pJDH119 (Hoheisel, 1989), and transformed into Escherichia coli JM101 using standard techniques (Sambrook et al., 1989). DNA isolation, purification in CsCl gradients, digestion with restriction enzymes and agarose gel electrophoresis were carried out using standard procedures (Sambrook et al., 1989).

DNA sequencing and computer analysis. Overlapping DNA fragments were sequenced on both strands using the dideoxynucleotide chain termination method (Sanger et al., 1977) following the procedures described in the protocol of the Circumvent Thermal Cycle Dideoxy DNA Sequencing Kit (New England Biolabs). For fragments which could not be subcloned further, oligonucleotides were synthesized for primers. Sequence analyses were performed using the UWGCG computer programs (Devereux et al., 1984) (version 7). The DNA sequence and the deduced amino acid sequence of the ORF were compared with sequences in the GenBank (release 75.0), EMBL (release 33.0), SWISSPROT (release 24.0) and NBRF databases, using the FASTA and TFASTA programs.

Results

AcMNPV fragment SstI-H (1.3 kbp) and fragment SstI–EcoRI (6 kbp), which form a contiguous segment on the right-hand end of fragment EcoRI-C located next to region hr3, were sequenced (Fig. 1). The approximate m.u. for this region are 47.0 to 52.5 (Vlak & Smith, 1982). The DNA sequence encompassed 7297 nucleotides, i.e. about 5% of the AcMNPV genome (Fig. 1). The locations of the most commonly used restriction sites are indicated in the nucleotide sequence (Fig. 2) and these sites were used to construct a restriction map for these enzymes (Fig. 1).

Eight ORFs were identified within the sequenced region and these are schematically shown in Fig. 1. The ORF annotations refer to the putative number of nucleotides. The predicted protein sequences for these eight ORFs are presented in Fig. 2. A partial ORF of 113 amino acids extends in the region upstream of m.u. 47. One of the ORFs has the same genomic DNA orientation as the polyhedrin gene (Vlak & Smith, 1982) and hence runs from left to right.

All but two ORFs (ORF 1227 and ORF 2541) are putative baculovirus genes since they contained putative transcriptional start and stop signals at appropriate locations up- and downstream from the ORF (Fig. 2, Table 1). Most early genes contain a promoter with the consensus sequence CAGT, whereas late AcMNPV gene promoters contain a motif (A/T/G)TAAG in which transcription initiates (see Blissard & Rohrmann, 1990, for review). Another promoter motif, CGTGC, found at the early start sites of the p143 gene (Lu & Carstens, 1992), the DNA polymerase gene (Tomalski et al., 1988), and at the early start site of p47 (Carstens et al., 1993) is also found upstream of several ORFs in the sequence presented here. All these motifs are indicated in Fig. 2, and their positions are also summarized in Table 1, together with the positions and sizes of all the putative
Fig. 2. For legend see opposite.
**Sequence of AcMNPV fragment EcoRI-C**

Fig. 2. Nucleotide and predicted protein sequences of 7297 bp of EcoRI-C. For ORFs running from left to right the amino acid sequence is placed above the DNA sequence. The amino acid sequences derived from the opposite strand are placed below the DNA sequence. Several restriction enzyme sites are underlined and named. Poly(A) signals and putative N-glycosylation sites are underlined. Potential consensus early and late transcription initiation sites are doubly underlined.

### Table 1. Summary of locations of ORFs, putative transcription and termination signals, sizes of the ORFs in nucleotides and amino acids, and the predicted sizes of the proteins

<table>
<thead>
<tr>
<th>ORF</th>
<th>Name</th>
<th>Position (nt)</th>
<th>Coding direction</th>
<th>Size (nt)</th>
<th>Transcription signals</th>
<th>Kozak consensus (ANNATGPu)</th>
<th>Poly(A) signal</th>
<th>Size (aa)</th>
<th>Mr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ORF &gt; 339</td>
<td>1-339</td>
<td>-</td>
<td>&gt; 339</td>
<td>GTAAG (-16)</td>
<td>A</td>
<td>T?</td>
<td>&gt; 113</td>
<td>13.2K</td>
</tr>
<tr>
<td>2</td>
<td>ORF 252</td>
<td>357-609</td>
<td>-</td>
<td>252</td>
<td>GTAAG (-14)</td>
<td>A</td>
<td>+241</td>
<td>84</td>
<td>9.5K</td>
</tr>
<tr>
<td>3</td>
<td>ORF 1137</td>
<td>627-1764</td>
<td>-</td>
<td>1137</td>
<td>GTAAG (-14)</td>
<td>A</td>
<td>+12</td>
<td>379</td>
<td>44.4K</td>
</tr>
<tr>
<td>4</td>
<td>ORF 327</td>
<td>1772-2099</td>
<td>-</td>
<td>327</td>
<td>GTAAG (-13)</td>
<td>A</td>
<td>-12</td>
<td>109</td>
<td>12.5K</td>
</tr>
<tr>
<td>5</td>
<td>ORF 312</td>
<td>2104-2416</td>
<td>-</td>
<td>312</td>
<td>GTAAG (-9)</td>
<td>A</td>
<td>+55</td>
<td>104</td>
<td>12.2K</td>
</tr>
<tr>
<td>6</td>
<td>ORF 1227 (gp41)</td>
<td>2421-3648</td>
<td>-</td>
<td>1227</td>
<td>GTAAG (-7)</td>
<td>A</td>
<td>-</td>
<td>409</td>
<td>45.4K</td>
</tr>
<tr>
<td>7</td>
<td>ORF 699</td>
<td>3640-4339</td>
<td>-</td>
<td>699</td>
<td>GTAAG (-9)</td>
<td>A</td>
<td>-18</td>
<td>233</td>
<td>26.9K</td>
</tr>
<tr>
<td>8</td>
<td>ORF 540</td>
<td>4190-4730</td>
<td>-</td>
<td>540</td>
<td>GTAAG (-9)</td>
<td>A</td>
<td>+50</td>
<td>180</td>
<td>20.0K</td>
</tr>
<tr>
<td>9</td>
<td>ORF 2541</td>
<td>4694-7234</td>
<td>+</td>
<td>2541</td>
<td>GTAAG (-52)</td>
<td>A</td>
<td>-</td>
<td>847</td>
<td>93.2K</td>
</tr>
</tbody>
</table>

ORFs. No consensus poly(A) signals were found immediately downstream from ORF 1227 or in the sequence immediately downstream from ORF 2541 or the sequence of hr3 (Guarino et al., 1986) which is located further downstream of ORF 2541.

Computer-assisted analysis of the sequence data indicated that there was no nucleotide or amino acid sequence homology with AcMNPV sequences deposited in data libraries, except for ORF 1227. This ORF, located on fragment SstII-M (m.u. 48.9 to 50), exhibited high sequence homology with the gp41 gene reported by Whitford & Faulkner (1992b), although this gene was much shorter (1062 bp) than ORF 1227, and had an opposite orientation. Recently, Whitford & Faulkner...
for a virion structural protein, p40, which may represent the homologue of AcMNPV gp41 (Fig. 3). Three separate regions (I, II and III) of 580, 896 and 227 bp, respectively, in the AcMNPV sequence showed homology with three separate similarly sized regions in the HbSNPV genome. The nucleotide sequence identity for each of these three regions is 64-1%, 65-4% and 56-8%, respectively. In both viruses these conserved regions are interspaced by non-conserved sequences (a and b) of different sizes (385 and 260 bp, respectively for AcMNPV, and 79 and 80 bp, respectively for HbSNPV) (Fig. 3).

The largest conserved region (II) corresponds with the C-terminal parts of the AcMNPV ORF 1227 product and HbSNPV p40. When the amino acid sequences of these proteins are compared a similarity of 79-1% and an identity of 62-4% was found over a stretch of approximately 300 amino acids (Fig. 4). In the C-terminal part of AcMNPV ORF 1227 a stretch of six amino acids, four of which are serines, is present (amino acids 164 to 169), which is absent from the HbSNPV p40 sequence. The major difference between AcMNPV ORF 1227 and HbSNPV p40 is found in the N-terminal sequence (Fig. 3, region b; Fig. 4). AcMNPV gp41 has an additional 86 amino acids at the N terminus, whereas p40 of HbSNPV has only 27 additional amino acids with almost no sequence homology with HbSNPV. This implies that the two ORFs encode polypeptides of different sizes, 409 amino acids for AcMNPV ORF 1227 and 322 amino acids for HbSNPV p40 (Fig. 4). The promoter regions of both ORFs (region III, Fig. 3) from position -227 (nucleotide residue 3875) to the translational start codon (ATG) show considerable sequence homology. A late ATAAG transcription signal is found at position -52 for AcMNPV and at position -49 for HbSNPV. In both viruses region III overlaps with the 3' end of AcMNPV ORF 699 and with an as yet unidentified ORF in HbSNPV showing 73-1% similarity and 46-2% identity over a stretch of 78 amino acids.

Region I of AcMNPV contains the complete ORF 327 and the 5' end of ORF 1137. In this region of the HbSNPV sequence an ORF of almost the same size (ORF 330) was found (Fig. 3). The predicted amino acid sequences of AcMNPV ORF 327 and HbSNPV ORF 330 showed a similarity of 64-4% and a identity of 44-6%. The homology between regions I ends near the ATG of both of these ORFs. The promoter regions for these ORFs show no homology (Fig. 3, region a), but share a late transcriptional motif (TTAAG), at position -94 for AcMNPV and at position -93 for HbSNPV. In this region AcMNPV has one unique ORF (Fig. 3, ORF 312). The 5' end of AcMNPV ORF 1137 and HbSNPV ORF 192 have a similarity of 73-3% and an identity of 61-7% at the amino acid level. The promoter regions of these ORFs show significant homology (64%), since they

![Fig. 3. Schematic presentation of regions of the AcMNPV and HbSNPV genomes with nucleotide sequence homology (hatched areas). For percentage nucleotide and amino sequence identities, see text.](image-url)

![Fig. 4. Comparison of amino acid sequences of AcMNPV ORF 1227 (gp41) (first line) and HbSNPV ORF 966 (p40) (second line).](image-url)
are part of ORF 327 and ORF 330, respectively. Again for both these ORFs a late TTAAG signal was found, at position -77 for AcMNPV and at position -60 for HszSNPV.

Computer-assisted analysis further indicated the absence of leucine zipper or zinc-finger motifs in the putative amino acid sequences of the ORFs presented here. Hydrophobicity analysis of these ORFs indicated that ORF 2541 contained a highly hydrophobic amino terminus (25 amino acid), which may represent a signal peptide. In most ORFs the N-terminal methionine is located in accordance with the Kozak translation initiation sequence ANNATGPu (Table 1) (Kozak, 1983).

**Discussion**

A contiguous sequence of 7297 nucleotides located on the EcoRI-C fragment of the AcMNPV genome has been determined (Fig. 1) and revealed eight different ORFs. The presence of transcriptional start and stop signals upstream and downstream of most of these ORFs suggest that they might be transcriptionally active, although this has yet to be proven experimentally. For five of the eight identified ORFs an early CAGT motif was found in the upstream sequence. To be transcriptionally active, this CAGT sequence is generally preceded by an upstream TATA box (Blissard *et al.*, 1992). However, no TATA boxes were found upstream of any of the CAGT sequences. Several transcripts, both early and late, have been mapped to the AcMNPV EcoRI-C fragment (Lübbert & Doerfler, 1984; Mainprize *et al.*, 1986). However, in view of the many ORFs on this fragment, a further allocation requires more detailed transcriptional analysis using ORF sequences as probes.

Only ORF 1227 is known to be transcriptionally active and this ORF has now been found to be homologous to the gp41 gene reported by Whitford & Faulkner (1992b). The nucleotide sequence of fragment SsrII-M revealed an additional G residue at position 2618 resulting in an ORF for gp41 which is 165 nucleotides longer than previously reported (Whitford & Faulkner, 1992b). Owing to this extra nucleotide, the last C-terminal 10 amino acids in the sequence of Whitford & Faulkner (1992b) are replaced by 65 additional amino acids. As a consequence the ORF encodes a polypeptide of 409 amino acids with a predicted size of 45.4K instead of 39.6K. The enlarged sequence of gp41 also reveals no hydrophobic C terminus, which supports the suggestion that gp41 is not an integral membrane protein (Whitford & Faulkner, 1992b). The gp41 is the major glycoprotein of polyhedra-derived virions with an Mr of 42K by PAGE (Stiles & Wood, 1983) and contains O-linked carbohydrates (Whitford & Faulkner, 1992a). According to our sequence data gp41 is thus in fact larger (> 45K). Antisera raised against the C-terminal part (Fig. 3, region b) of gp41 using an oligopeptide as antigen might be used to confirm the existence of the ORF 1227 product in polyhedra-derived virions.

An alternative explanation for the difference in gp41 sequence could be that another strain of AcMNPV (HR3) was used in the study of Whitford & Faulkner (1992b). The extra C-terminal 65 amino acids of ORF 1227 show homology to the C terminus of HszSNPV p40, an observation which supports the sequence of ORF 1227. However, the last 22 amino acids of ORF 1227 are not present in p40. A difference in size is also found at the N termini of both ORFs. AcMNPV ORF 1227 has 86 amino acids at the N terminus which show almost no homology with the only 27 amino acids at the N terminus of HszSNPV (Fig. 3, region b; Fig. 4). The high degree of homology in this region between AcMNPV and HszSNPV suggests a conserved sequence organization in MNPs and SNPs.

A 2.1 kb transcript encoding the gp41 protein was detected by Whitford & Faulkner (1992b). The size of this transcript corresponds with the observation that no poly(A) signal was found immediately downstream of ORF 1227 (gp41), as this would have given a transcript of about 1.5 kb. When the sequence downstream of the gp41 ORF was searched for transcription termination signals in agreement with the size of the actual transcript, two are found at residues 1717 and 1784. This suggests that in AcMNPV this transcript most likely proceeds through two other ORFs, ORF 312 and ORF 327, and terminates at or near one of the poly(A) signals in or behind the sequence of ORF 327. This may suggest that the transcript is tricistronic. Several other polycistronic transcripts have been mapped to the AcMNPV genome (see Kool & Vlak, 1993, for review). The major p39 capsid gene and the downstream CG30 polypeptide gene are present on a dicistronic mRNA (Thiem & Miller, 1989). It is not clear whether both genes are translated from the same message, since CG30 is expressed early and p39 early (see Kool & Vlak, 1993, for review). Using baculoviruses as an expression vector Haseman & Capra (1990) showed that downstream ORFs on a bicistronic messenger RNA are not translated. The short intergenic regions between ORF 1227, ORF 312 and ORF 327, and the compatible size of these three ORFs with the transcript of 2.1 kb (Whitford & Faulkner, 1992b) supports the hypothesis of a tricistronic message. In HszSNPV ORF 312 is missing (Ma *et al.*, 1993).

Regions of the EcoRI-C fragment that were maintained in defective genomes (Kool *et al.*, 1993a) map predominantly within ORF 2541. Further analysis of this region is now facilitated by the availability of the nucleotide sequence. This analysis may reveal cis-acting
elements involved in the generation and/or encapsidation of the defective viruses.

We thank Dr Peter Faulkner, Kingston, Canada and Dr John Kuzio, Oxford, U.K. for reviewing this manuscript and communicating sequence data. This research was supported in part by the Dutch Programme Committee for Agricultural Biotechnology and the Netherlands Foundation for Chemical Research (SON) with financial aid from the Netherlands Organization for Scientific Research (NWO).

References


