Nucleotide sequence of the polyhedrin gene region of Helicoverpa zea single nucleocapsid nuclear polyhedrosis virus: placement of the virus in lepidopteran nuclear polyhedrosis virus group II

P. Cowan, D. Bulach, K. Goodge, A. Robertson and D. E. Tribe*

Department of Microbiology, University of Melbourne, Parkville, Victoria 3052, Australia

The polyhedrin gene (polh) of Helicoverpa zea single nucleocapsid nuclear polyhedrosis virus (HzSNPV) was identified and shown by sequence analysis of the EcoRI I genomic fragment to encode a 246 amino acid polypeptide that has greater than 80% sequence identity to known polyhedrins. It is preceded by an AT-rich region containing the conserved late promoter motif TAAG, which was identified as a transcription start point. Downstream of polh there were several similarities in genome arrangement to other nuclear polyhedrosis viruses (NPVs). These include open reading frame (ORF) 8, immediately downstream of polh, encoding a 412 amino acid protein with multiple tandem proline residues, which is homologous to ORF8 (ORF1629) of Autographa californica multiple nucleocapsid NPV. Phylogenetic analysis of the polh gene region shows that HzSNPV is a member of the previously described lepidopteran NPV group II and that it is most closely related to polh of the NPVs of Malacosoma nuesta, Spodoptera littoralis, Orgyia pseudotsugata (single nucleocapsid-type virus) and Buzura supressaria.

Baculoviruses are insect viruses with large circular DNA genomes, and include the genera Nuclear polyhedrosis virus (NPV) and Granulosis virus (GV). NPVs form polyhedral occlusion bodies and are found in several arthropod orders, including Lepidoptera and Hymenoptera, whereas GV forms smaller, ellipsoidal cytoplasmic occlusions and are found only in the Lepidoptera (Rohrmann, 1992). Understanding of phylogenetic relationships among baculoviruses is highly relevant to the future development of biopesticides and to understanding of virus host range, but this information has emerged slowly (Zanotto et al., 1993). Gene sequence data relating to occlusion body proteins (polyhedrin, granulin) are proving to be an important tool for discerning phylogenetic relationships, recently allowing lepidopteran NPV evolution to be divided into two distinct branches (Zanotto et al., 1993), namely group I and group II. However, many NPVs remain unclassified. In this report we give the sequence of the Helicoverpa (formally Heliothis) zea single nucleocapsid NPV (HzSNPV) polyhedrin gene (polh) region, and present evidence for inclusion of HzSNPV in group II.

Lepidoptera is the most recently evolved insect order (Reik, 1970), diversifying in the last 40 to 60 million years. Studies of occlusion body proteins suggest that the numerous different lepidopteran NPVs have evolved from a lepidopteran NPV rather than cross-infecting from other orders of arthropods, and that the GV evolved from a lepidopteran NPV early in this adaptive radiation (Rohrmann et al., 1981; Vlak & Rohrmann, 1985; Rohrmann, 1992; Zanotto et al., 1993). Previous emphasis on virus morphotype (for example Rohrmann et al., 1981; Vlak & Rohrmann, 1985) has been a source of confusion in classification. It is now held to be of uncertain value in establishing phylogenies (Zanotto et al., 1993), as multiple nucleocapsid GVVs have been discovered (Falcon & Hess, 1985) and single nucleocapsid taxa branch from within groups of multiple nucleocapsid NPV species in phylogenetic trees.

The baculoviruses used in this study are shown in Table 1. Sequence analyses were performed using (i) PHYLIP version 3.5 Protpars, DNApars, DNAml 3.5c, Seqboot and Consense programs (Felsenstein, 1993), (ii) Clustal V multiple alignment and neighbor-joining (distance) algorithms (Higgins et al., 1992; Saitou & Nei, 1987) and (iii), for sequence similarity assessment, Rdf2,
Table 1. Baculoviruses used in this study

<table>
<thead>
<tr>
<th>Species of origin</th>
<th>Single (S) or multiple (M) of nucleocapsid</th>
<th>Genus of virus</th>
<th>Abbreviation</th>
<th>Group†</th>
<th>Source of sequence data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helicoverpa zea</td>
<td>S</td>
<td>NPV</td>
<td>HzSNPV</td>
<td></td>
<td>This paper</td>
</tr>
<tr>
<td>Trichoplusia ni</td>
<td></td>
<td>GV</td>
<td>PbGV</td>
<td></td>
<td>Chakerian et al. (1985)</td>
</tr>
<tr>
<td>Agrotis segetum</td>
<td></td>
<td>GV</td>
<td>AsGV</td>
<td>I</td>
<td>Kozlov et al. (1992)</td>
</tr>
<tr>
<td>Pieris brassicae</td>
<td></td>
<td>GV</td>
<td>AsNPV</td>
<td>I</td>
<td>Kozlov et al. (1992)</td>
</tr>
<tr>
<td>Anticarsia gemmatalis</td>
<td>M</td>
<td>NPV</td>
<td>AgMNPV</td>
<td>I</td>
<td>Zanotto et al. (1992)</td>
</tr>
<tr>
<td>Autoaphaga california</td>
<td>M</td>
<td>NPV</td>
<td>AcMNPV</td>
<td>I</td>
<td>Hooft Van Iddekinge et al. (1983)</td>
</tr>
<tr>
<td>Bombyx mori</td>
<td></td>
<td>NPV</td>
<td>BmNPV</td>
<td>I</td>
<td>Iatrou et al. (1985)</td>
</tr>
<tr>
<td>Buzura supressaria</td>
<td>S</td>
<td>NPV</td>
<td>BsSNPV</td>
<td>I</td>
<td>Hu et al. (1993)</td>
</tr>
<tr>
<td>Hyphantria cunea</td>
<td></td>
<td>NPV</td>
<td>HcNPV</td>
<td>S</td>
<td>S. Isayama; D15473‡</td>
</tr>
<tr>
<td>Lymantria dispar</td>
<td></td>
<td>NPV</td>
<td>LdMNPV</td>
<td>I</td>
<td>Smith et al. (1988)</td>
</tr>
<tr>
<td>Malacosoma nustria</td>
<td></td>
<td>NPV</td>
<td>MnNPV</td>
<td>I</td>
<td>Petenko et al. (1994)</td>
</tr>
<tr>
<td>Mammestra brassicae</td>
<td>M</td>
<td>NPV</td>
<td>MsMNPVpro</td>
<td>II</td>
<td>E. Koslov &amp; A. Solomko (unpublished)</td>
</tr>
<tr>
<td>Orgysia pseudotsugata</td>
<td>M</td>
<td>NPV</td>
<td>OpMNPV</td>
<td>I</td>
<td>Leisy et al. (1986b)</td>
</tr>
<tr>
<td>Panolis flammea</td>
<td></td>
<td>NPV</td>
<td>PIMNPV</td>
<td>II</td>
<td>Leisy et al. (1986a)</td>
</tr>
<tr>
<td>Porthetria dispar</td>
<td></td>
<td>NPV</td>
<td>PaMNPV</td>
<td>I</td>
<td>Oakley et al. (1989)</td>
</tr>
<tr>
<td>Spodoptera frugiperda</td>
<td>M</td>
<td>NPV</td>
<td>SiMNPV</td>
<td>II</td>
<td>Kozlov et al. (1981)</td>
</tr>
<tr>
<td>S. exigua</td>
<td></td>
<td>NPV</td>
<td>SeMNPV</td>
<td>II</td>
<td>Gonzalez et al. (1989)</td>
</tr>
<tr>
<td>S. littoris</td>
<td></td>
<td>NPV</td>
<td>SiMNPVpro</td>
<td>II</td>
<td>van Strien et al. (1992)</td>
</tr>
</tbody>
</table>

* Lepidopteran unless otherwise stated.
† Group of lepidopteran NPVs, defined by Zanotto et al. (1993).
‡ EMBL database accession number.

The polh gene has been mapped to the EcoRI I fragment of HzSNPV derived from the same source as our isolate (Knell & Summers, 1984). A restriction map of the region of the HzSNPV genome from 95 to 128 map units (m.u.) was assembled by establishing the contiguous nature of inserts in the plasmid clones. These were EcoRI E, EcoRI I and HindIII E fragments as shown in Fig. 1. These designations accord with fragment nomenclature used previously (Knell & Summers, 1984). That the EcoRI I and HindIII E fragments overlapped was established by restriction digest analysis of the putative overlap and verified by dideoxynucleotide sequencing of both clones. Contiguity of EcoRI E and EcoRI I was established by PCR and DNA sequencing across the junction using the viral DNA template (data not shown).

DNA sequencing of the EcoRI I fragment confirmed that the polh gene straddles the HindIII site at m.u. 1-8. Fig. 1 shows the orientation and relative position of polh with respect to flanking markers and restriction sites and Fig. 2 shows the sequence data on which identification of the polh gene and the adjacent open reading frame (ORF) 8 was based. Fig. 3 shows an alignment of the
I predicted ORF8 protein with the corresponding protein of AcMNPV (Possee et al., 1991).

The restriction map of our isolate, both upstream and downstream of polh, largely agrees with that previously reported for strain Elkar (Knell & Summers, 1984), and is quite different from that reported for isolate HZS-15, which is known to have undergone DNA rearrangement (Corsaro & Fraser, 1987; Ma et al., 1993). The HindIII fragment containing ORF8 and the kinase gene in Fig. 1 resembles the HindIII L fragment defined by Knell & Summers (1984), and is unlike a similarly mapped HindIII I fragment of isolate HZS-15. In Fig. 1, as in the map reported by Knell & Summers (1984), there is a missing HindIII site upstream of polh and a missing BamHI site downstream, both being present in the HZS-15 isolate (Corsaro & Fraser, 1987). The distinctive pattern of BamHI sites present in the HindIII E fragment (Fig. 1), confirmed by sequence analysis of both strands, corresponds to previously identified BamHI sites at m.u. 12 (Knell & Summers, 1984), but in isolate HZS-15 similar sites are at m.u. 36 to 39.

It has been established (Ma et al., 1993) that in isolate HZS-15 the p40 structural protein gene terminates upstream of polh, near m.u. 97, and thus mapping of p40 is relevant to our polh map. In our strain we have identified the p40 gene in the HindIII K and EcoRI H fragments (Fig. 1). Our DNA sequence analysis was almost identical at the translated amino acid level to the previously reported p40 sequence (248 out of 249 residues identical). We confirmed that the map of this region was largely as reported (Knell & Summers, 1984) and identified an additional small EcoRI X fragment in the DNA polymerase ORF (Fig. 1). Furthermore, direct sequence analysis of the EcoRI M and HindIII K fragments showed that there was no polyhedrin-related sequence for 6.7 kbp downstream from the p40 terminus (data not shown). In this 6-7 kb region we identified a 1020 amino acid ORF encoding a DNA polymerase-related protein (see Table 2). This analysis rules out a location for p40 near polh in our strain but is consistent with placement of the DNA polymerase-related ORF and p40 on the reported map (Knell & Summers, 1984) near m.u. 50. We interpret the agreement of our results with a previous finding (Knell & Summers, 1984) and the radically different maps proposed by other investigators with the rearranged strain (Corsaro & Fraser, 1987; Ma et al., 1993) as evidence that genomic rearrangements are more extensive than previously suspected.

The HzSNPV polh gene (Fig. 2) was identified as an ORF predicted to code for a protein of 246 residues.
CAAGATTTTGAAGAGTTAAATGTTGTTTACTCGACAATAACATACCAATGTATACGGCTTTGTGTGATGATGTGTTTGA
AAAGCAAACCAATTATACAAGATATAGAAATATGTAATGAACAAAATATGCGTTGAAGGAGCGTACGTGCCATTCTCAAGAAG
AAATTTTGAATATCAAATCTTTTTGCAAGAATATGAAGATTTCTGTCGTGTGAAAATTTGTAATAAAACTAAATAA
AACCTTATAATATA~TATTA~CATACACTTTTATTTCTA~TAAGTATTTTTTTCCTATTGTTCAAGATTGTGA~

Fig. 2. Nucleotide sequence and predicted amino acid sequence of the HzSNPV polh gene. Two major ORFs, designated polh and ORF8 respectively, were found in the sequence. The late gene conserved promoter motif is underlined and poII-related motifs overlined. A putative poly(A) signal is in bold. Transcription start sites, identified by primer extension analysis (data not shown), are indicated by asterisks above the sequence. Double lines show the 51 nucleotide conserved promoter region used for phylogenetic analysis.
proteins by comparison of translated ORFs 

et al.

Slavicek

Table

protein of HzSNPV with regions A and B of the ORF8 protein of AcMNPV. Regions of tandemly repeated proline residues are present in both proteins. Tests for the significance of sequence similarity of regions A and B are given in Table 2.

<table>
<thead>
<tr>
<th>HSNPV loci</th>
<th>Map position in m.u.*</th>
<th>Sequence data used</th>
<th>Gene for comparison†</th>
<th>Optimized Rdf2 score in S.D. above mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORF8</td>
<td>2.5</td>
<td>Region A of Fig. 3</td>
<td>AcMNPV ORF8</td>
<td>2.7 (25%)‡</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Region B of Fig. 3</td>
<td>AcMNPV ORF8</td>
<td>27 (23%)</td>
</tr>
<tr>
<td>Protein kinase-related ORF</td>
<td>3.7</td>
<td>69 amino acid residues</td>
<td>LdMNPV protein kinase ORF</td>
<td>33 (55%)</td>
</tr>
<tr>
<td>ie1</td>
<td>11-12</td>
<td>649 amino acid residues</td>
<td>AcMNPV ie1</td>
<td>43 (32%)</td>
</tr>
<tr>
<td>DNA polymerase</td>
<td>(50)</td>
<td>1020 amino acid residues</td>
<td>LdMNPV DNA polymerase</td>
<td>252 (52%)</td>
</tr>
</tbody>
</table>

* Position relative to a unique XhoI site at 0 m.u.
† References: AcMNPV ORF8, Possee et al. (1991); LdMNPV protein kinase ORF, Bischoff & Slavicek (1994); AcMNPV ie1, Guarino & Summers (1987); LdMNPV DNA polymerase, Bjornson et al. (1992).
‡ Degree of sequence identity found using ALIGN.

Fig. 3. Alignment of the 412 amino acid residues of the predicted ORF8 protein of HzSNPV with regions A and B of the ORF8 protein of AcMNPV. Regions of tandemly repeated proline residues are present in both proteins. Tests for the significance of sequence similarity of regions A and B are given in Table 2.
BsSNPV, OpSNPV, SiMNPV and MnNPV promoters. This region has motifs similar to the polIII class of promoters previously discussed (Zanotto et al., 1992). A potential poly(A) signal sequence was also found downstream of polh, within ORF8. The stop codons of the convergent polh and ORF8 reading frames overlap. A similar sequence to part of this region has been reported by another laboratory (Fraser et al., 1989). The sequence differs in predicting a polyhedrin protein with one residue deleted and another changed.

An alignment of the putative ORF8 protein with the ORF8 (ORF1629) protein of AcMNPV is shown in Fig. 3. ORF8 of AcMNPV appears to code for an essential protein with a function that is presently not understood (Possee et al., 1991; Pham & Sivasubramanian, 1992). A distinctive feature of both proteins is the presence of multiple proline residues about one third of the way through the sequence. In the AcMNPV protein there are two polyproline tracts separated by 70 amino residues, but in HzSNPV these regions are almost adjacent (Fig. 3). Apart from the polyproline regions, the proteins are not very similar, but region A and region B (Fig. 3) show two separate regions of detectable homology identified by dot-matrix analysis. Results of tests on the significance of these distant homologies are given in Table 2. The match in region A is barely significant, but for region B it is highly significant. Polyproline tracts are also found in the ORF downstream of polh in MbMNPV (Cameron & Possee, 1989) but other sequence similarities to ORF8 of either HzSNPV or AcMNPV are not obvious. The ORF8 homologue of BmNPV is highly similar to that of AcMNPV (Iatrou et al., 1985).

Downstream of ORF8 we detected sequence homology to the protein kinase genes of AcMNPV and LdMNPV, and at m.u. 11 we located the immediate early 1 (iel) protein trans-activator gene (see Fig. 1, Table 2). The genomic location of the protein kinase gene of AcMNPV (ORF9) is the same as the HzSNPV protein kinase-related sequence (Kool & Vlak, 1993; Possee et al., 1991) and iel of HzSNPV has a similar placement to iel in OpMNPV (Theilmann & Stewart, 1991). In light of the inversion of polh proposed for AcMNPV and AgMNPV (Zanotto et al., 1993) the apparent difference in iel location from that of AcMNPV (Guarino & Summers, 1987) should be interpreted cautiously. These homologies between genes flanking ORF8 confirm our conclusion that ORF8 of HzSNPV and AcMNPV are homologous, and show that HzSNPV gene order in the region 0 to 12 m.u. is preserved in some other NPVs.

Phylogenetic relationships between the polh gene of HzSNPV and those of other lepidopteran baculoviruses, with particular emphasis on the relationship with NPV groups I and II, were examined using distance, parsimony and maximum likelihood methods (Swofford & Olsen, 1990). First, occlusion protein sequence data were used to establish homologous positions by sequence alignment. These alignments were (i) a 24 taxa protein data set, (ii) a 15 taxa data set of polyhedrin DNA coding regions and (iii) a promoter DNA data set corresponding to the previously mentioned 51 nucleotide promoter region (Fig. 2) that required no gaps for alignment. The alignment and choice of gaps for the protein and coding DNA, checked using Clustal V software, were essentially as reported (Zanotto et al., 1993) with additional sequences added (HzSNPV, BsSNPV, MnNPV, MnNPVpro, SiMNPV, AsNPV and AsGV) and a different, complete sequence for AsNPV. Polyhedrin of SiMNPV was observed to have an insertion of three residues near the N terminus. Identification of this insertion as N-terminal MYTNRYS- was made by investigating the relatedness of polypeptide sequences with N-termini omitted. Using either distance or protein parsimony, SiMNPV consistently nested within the main group II branch, confirming that the N-terminal MYTNRYS- motif of SiMNPV is appropriately aligned with a conserved MYTNRYS- present in group II taxa.

Distance and parsimony evaluation (data not shown) of the three data sets (protein, promoter, DNA coding) allowed us to identify the root of the trees generated from DNA data sets. For example, for the tree shown in Fig. 4, which is based on pooled promoter and DNA coding sequences (discussed below), consideration of the position of the GV and hymenopteran NsSNPV branches indicated that the root of the tree lies between node I and node II (Fig. 4) and that the LdMNPV taxon branches near this root. The various trees obtained from both distance and parsimony examination of promoter, coding DNA, or protein data sets separately consistently had a branch comprising the same set of taxa as those branching from node II in Fig. 4. BsSNPV and OpSNPV were most closely similar in polypeptide sequences to HzSNPV (91.7% and 91.2% identity, respectively). There was also consistent support for close relationships between SiMNPV, MnNPV and HzSNPV, and many similarities with the previously reported trees for baculoviruses (Zanotto et al., 1993). When reliability of the trees obtained from the separate data by either distance or parsimony methods was examined (data not shown) by bootstrap analysis (Felsenstein, 1985) only moderate support for these inferences was obtained (for example, 50% support for the clade defined by node II of Fig. 4).

Pooling the promoter and polyhedrin-coding DNA data as in Fig. 4 allowed more convincing inferences to be drawn. Results of maximum likelihood analysis (program DNAml version 3.5c; Felsenstein, 1985; using the global rearrangements option) are presented since they offer confidence interval estimates on branch
lengths, and they may be superior tree-making algorithms (Saitou & Imanishi, 1989). This approach uses all of the sites in the sequence and allows different rates of evolution at different sites. With LdMNPV as an outgroup, the tree in Fig. 4 suggests that group II includes a wider range of taxa than previously identified. It also suggests that group II is divided into two distinct clades, indicated on the tree at nodes A and B. The lengths of the branches in the tree are scaled in proportion to the expected number of nucleotide substitutions, and we note that although the branch leading to node IIB is shorter than that to I, it is marginally longer than that to node IIA. The reliability of this phylogeny was supported by confidence estimates on the branch lengths, which are all greater than zero (based on the confidence interval $P < 0.01$). Different choices of parameters for the maximum likelihood calculation (for example three classes of site, in the ratio of 1:1:1, changing at rates in the ratio of 2.5:1:20 as reported for nucleotide change in NPV; Zanotto et al., 1993) gave essentially the same tree. A phylogeny as in Fig. 4 is also supported by a different analysis of the pooled data using DNA parsimony. For this analysis (data not shown) the third base of each codon was excluded to avoid bias. Bootstrapping gave 75% support for node II (this may be an underestimate of the support as it was not practical to exclude uninformative sites from bootstrapping). Further support for the HzSNPV, SIMNPV and MnPV clade defined by node IIC was obtained from both parsimony- and distance algorithm-based trees generated from the protein sequence data set (data not shown) which included a second MnPV sequence for which DNA data were unavailable.

These evaluations all support the inclusion of HzSNPV as a member of group II and imply that evolution of group II may have proceeded in two lines, corresponding to nodes IIA and IIB in Fig. 4. A different placement of the OpSNPV–BsSNPV branch in protein trees as compared to DNA-based trees (Zanotto et al., 1993; data not shown) indicates that its inclusion in the IIB clade is provisional. It has been suggested that the group II taxon SeMNPV poly(A) mRNA is not polyadenylated and that other group II taxa, unlike group I viruses, lack obvious poly(A) signals (Zanotto et al., 1993). The presence of putative poly(A) signals in HzSNPV (Fig. 2), OpSNPV (Leisy et al., 1986a) and BsSNPV (Hu et al., 1993) may distinguish them from other group II taxa. The existence of an independent lineage of single nucleocapsid viruses in group II is still an open question. Continued molecular analysis of the conserved DNA polymerase, protein kinase, p40 and ile genes should prove invaluable in elucidating this and other relationships among NPVs.

HzSNPV and H. armigera were kindly provided by Dr. Ron Atkinson, University of Southern Queensland. E. Koslov and A. Solomko are thanked for access to their unpublished results and the University of Melbourne for financial support.

**References**


(Received 5 January 1994; Accepted 27 June 1994)