Wang, W., et al. (2001). Identification, sequence analysis, and phylogeny of the immediate early gene 1 of the trichoplusia ni single nucleocapsid polyhedrosis virus. Virus Genes, 23(1): 53-62

Identification, Sequence Analysis, and Phylogeny of the Immediate Early Gene 1 of the *Trichoplusia ni* Single Nucleocapsid Polyhedrosis Virus

Weizhou Wang, Neil Leat, Burtram Fielding and Sean Davison

Abstract. Substantial research has been conducted on the immediate early 1 (*ie-1*) genes from the prototype baculovirus *Autographa californica* multicapsid nuclear polyhedrosis virus (AcMNPV) and the *Orgyia pseudotsugata* multicapsid nuclear polyhedrosis virus (OpMNPV). In both cases *ie-1* gene products have been implicated in transcriptional activation and repression. In this study an *ie-1* homolog was identified from *Trichoplusia ni* single nucleocapsid polyhedrosis virus (TniSNPV). Nucleotide sequence analysis indicated that the TniSNPV *ie-1* gene consists of a 2217 nucleotide open reading frame (ORF), encoding a protein with a molecular mass of 84.464 kDa.

This represents the largest baculovirus *ie-1* gene characterised to date. Of the seven *ie-1* homologs identified to date, the TniSNPV *ie-1* shared most sequence similarity with the *ie-1* gene of *Spodoptera exigua* MNPV (SeMNPV) (41%). At the nucleotide level, expected TATA and CAGT motifs were found to precede each *ie-1* ORF. At the protein level, it was confirmed that the N-termini are poorly conserved, but share the characteristic of having a high proportion of acidic amino acids. In addition it was found that N-terminal regions significantly matched the SET domain in the Swiss-Prot prosite database. The C-terminal regions of the deduced IE-1 sequences were found to be substantially more conserved than the N-termini. Several conserved motifs were identified in the C-terminal sequences. A phylogenetic tree of nine baculovirus IE-1 proteins was constructed using maximum parsimony analysis. The phylogenetic estimation of the *ie-1* genes shows that TniSNPV is a member of the previously described lepidopteran NPV group II and it is most closely related to SeMNPV.

Introduction

Baculoviruses are a diverse group of insect viruses with circular double strandedDNAgenomes ranging in size between 88 and 160 kb [1]. Among nuclear polyhedrosis viruses, the well-characterized *Autographa californica* multicapsid nuclear polyhedrosis virus (AcMNPV) encodes approximately 150 genes [2]. The expression of baculovirus genes occurs in a transcriptional cascade occuring in essentially four phases: immediate early, delayed early, late and very late [3].

Immediate early (IE) and delayed early (DE) genes are expressed before the onset of viral DNA replication and their transcription is dependent on host RNA polymerase II [4]. Late and very late genes are expressed after or concurrently with viral DNA replication, and their transcription is facilitated by a virus-specificRNA polymerase [1,3]. The products of the immediate early 0 and 1 genes (*ie-0*) and (*ie-1*) play a central role in the regulation of viral gene expression.

The most extensively studied *ie-0* and *ie-1* genes are those of the AcMNPV. The *ie-0* transcript is a spliced product of the *ie-1* transcript [5]. The protein product of the *ie-0* transcript differs from that of *ie-1* by the addition of 54 N-terminal amino acids. Both *ie-0* and *ie-1* are expressed shortly after infection, however, translation of *ie-0* ceases in the early stages of infection while expression of *ie-1* persists throughout infection [5].

The AcMNPV *ie-1* gene product mediates three forms of transcriptional regulation. Firstly it is involved in transcriptional induction of promoters associated with homologous repeats (*hr*), present on theAcMNPV genome. This involves binding *ie-1* dimers to 28 bp palindromes present in *hr* elements and transcriptional activation of associated promoter [6]. This has been demonstrated to occur at the 39K promoters [5]. The second form of regulation involves transcriptional activation in an *hr* independent manner. This has been also demonstrated to occur at the 39K promoters [7]. Finally *ie-1* is capable of negatively regulating transcription. This has been demonstrated to occur at the *ie-0*, *ie-2*, and *pe-38* promoters [5,8,9].

Examination of amino acid sequences at the N-terminus of the IE-1 proteins of AcMNPV and *Orgyia pseudotsugata* muticapsid nuclear polyhedrosis virus (OpMNPV) revealed that they are poorly conserved and rich in acidic residues, while C-terminal residues are more conserved [10]. Attempts have been made to map functional domains within the IE-1 protein of AcMNPV. Initially this involved deletions of the N- and C-terminal domains demonstrating that N-terminal deletions affected transcriptional activation while C-terminal deletions prevented DNA binding [11].

In an attempt to map domains involved in transactivation, N-terminal residues from IE-1 were fused to the DNA binding domain of the GAL4 transactivator [12]. The resulting fusion proteins activated transcription at promoters linked to GAL4 DNA binding sites. This approach mapped the minimum region of IE-1 required for transcriptional transactivation to N-terminal residues 8–118. A similar study using the LacI DNA binding domain mapped IE-1 domains involved in transcriptional transactivation to the first 266 N-terminal amino acid residues [13]. Attempts to map domains involved in DNA binding and oligomerization involved insertional mutagenesis of C-terminal residues [12]. This demonstrated that insertions at residues 425 and 553 disrupted DNA binding and abolished IE-1 oligomerization respectively [12].

Baculoviruses have been previously taxonomically subdivided into two genera, nuclear polyhedrosis virus (NPV) and granulosis virus (GV) [14]. For understanding of phylogenetic relationship among baculoviruses, gene sequence data relating to occulsion body proteins (polyhedrin and granulin) were used to resolve lepidopteran NPV evolution and to classify the NPV into two distinct branches, group I and group II [15].

Further studies have defined the group II into subclades II-A, -B and -C [16]. Theseworks have provided useful data for understanding of virus host range and the further development of biopesticides. Our previous study of *Trichoplusia ni* single nucleocapsid polyhedrosis virus (TniSNPV) and other polyhedrin homologs suggested that TniSNPV is more closely related to group II NPVs than group I [17]. However, without the estimation of a phylogenetic tree the evidence for inclusion of TniSNPV in group II cannot be conclusive, especially for establishing a phylogenetic relationship of deep branches.

In the present study an *ie-1* homolog was identified from TniSNPV. Here we present an analysis of the TniSNPV *ie-1* nucleotide and deduced protein sequences. All eight currently available NPV *ie-1* nucleotide and deduced protein sequences are compared to identify common features. In addition, the phylogenetic relationship of TniSNPV in the baculovirus classification was also established.

Materials and Methods

Insects and Virus

The original virus isolate was prepared from diseased *Trichoplusia ni* (Noctuidae: Lepidoptera) larvae collected from the Eastern Cape, South Africa [17]. This virus was routinely propagated in third instar *Trichoplusia ni* larva. Larvae were fed on an artificial lepidoperan diet and reared at 26◦C and 65% humidity on a 12 h day/night cycle. Virus was purified from larvae as previously described [18].

DNA Extraction and Manipulation

TniSNPV DNA was isolated from the occluded form of the virus [17]. The viral genome was digested with *Eco*R1 and the fragments ligated into pSKBluescript (Stratagene). As part of a preliminary attempt to map the genome the ends of the *Eco*R1 library were sequenced. Subsequent genome analysis led to the identification of the TniSNPV homolog. The *ie-1* gene was found to be truncated with its 5' and 3' ends on a 11 kb and a 2.3 kb *Eco*RI fragment respectively. Appropriate templates were prepared for nucleotide sequencing by exonuclease III digestion [19]. Sequencing was conducted using the Sequitherm kit (Epicentre Technologies) using CY-5 labeled primers. Nucleotide sequence was resolved on an Alfexpress automated DNA sequencer (Pharmacia). Sequence was obtained in both the sense and anti-sense directions before the final sequence was confirmed (Fig. 1).

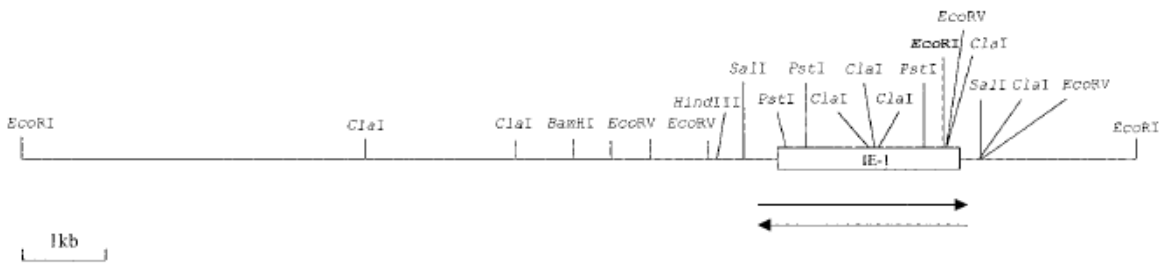

Fig. 1. The location of the *ie-1* gene within an 11 kb *Eco*RI fragment and an adjoining 2.3 kb *Eco*RI fragment. An *Eco*RI restriction site between the 11 and 2.3 kb fragments is in bold. The sequencing directions are represented by arrows.

Computer Analysis

Nucleotide and amino acid sequence manipulation was carried out using the University of Wisconsin, Genetics Computer Group (GCG) sequence analysis package. The BLAST algorithm of Altschul et al. 1990 [20] was used to compare sequences generated in this study with entries in non-redundant nucleotide and protein sequences databases accessed by the National Center for Biotechnology Information (NCBI). Multiple sequence alignments were conducted using the ClustalW program of Thompson et al. 1994 [21]. The alignment was used as the input to construct the phylogenetic tree with heuristic search settings of PAUP 4.0b4a [22]. Under maximum parsimony, we also tested for the robustness of the data sets using the bootstrap method [23] implemented in PAUP 4.0b4a. GenDoc software was used for similarity shading and scoring among the aligned sequences.

Results and Discussion

At present *ie-1* homologs have been isolated from seven baculoviruses. While detailed studies have been conducted on the *ie-1* genes of AcMNPV and OpMNPV, the remaining *ie-1* genes have not been extensively characterized. In the present study, a gene encoding a TniSNPV *ie-1* homolog was isolated. We present an analysis of the TniSNPV *ie-1* sequence and a detailed comparison of *ie-1* homologs.

Analysis of the nucleotide sequence of the TniSNPV *ie-1* gene, revealed an open reading frame (ORF) of 2217 nucleotides, encoding a protein of 739 amino acids, with a molecular mass of 84.464 kDa (Fig. 2). If the predicted start and stop codons are correct it would mean that the TniSNPV *ie-1* ORF is slightly larger than that reported for previously studied *ie-1* genes. Other baculovirus *ie-1* ORFs range from 1680 to 2142 nucleotides for OpMNPV and *Spodoptera exiqua* MNPV (SeMNPV) species respectively. The nucleotide sequence of the TniSNPV *ie-1* gene is relatively A/T rich. This is similar to the nucleotide compositions of most *ie-1* genes, but differs from that of OpMNPV *ie-1* and *Lymantria dispar* MNPV (LdMNPV) *ie-1* which are G/C rich . The most favored codons of the TniSNPV *ie-1* ORF are AAT and AAC (asparagine) and AAA (lysine). The least favored ones are AGG and CGG (arginine), TGG (tryptophan), GCG (alanine), TGT (cysteine), CTT (leucine), CAC (histidine) and GTA (valine). Total number of negatively charged residues (aspartic acid and glutamic acid) and positively charged residues (arginine and lysine) are 87 and 101, respectively.

Several *cis*-acting regulatory elements such as the TATA element and the tetranucleotide CAGT are conserved in the promoter regions of immediate early and early genes [3]. Equivalents of these elements were previously identified in the promoters of AcMNPV and OpMNPV [24]. The TATA element enhances expression of the associated promoter while the CAGT element was shown to act as an initiator element, determining the position of transcription initiation in the case of the AcMNPV *ie-1* [24]. As expected, a comparison of the promoter regions of all of the available *ie-1* genes revealed that the TATA and CAGT elements are highly conserved in their promoters (Fig. 3). Potential TATA motifs are present in every case. Potential CAGT motifs are almost completely conserved, with *Helicoverpa Zea* SNPV (HzSNPV) being the only example where the motifs are not completely conserved.ACAGAis present at−33 upstream of the translation

initiation codon of HzSNPV *ie-1*. A single polyadenylation signal incorporating the TAA termination codon was found at the 3_ end of the TniSNPV *ie-1* ORF (Fig. 2).

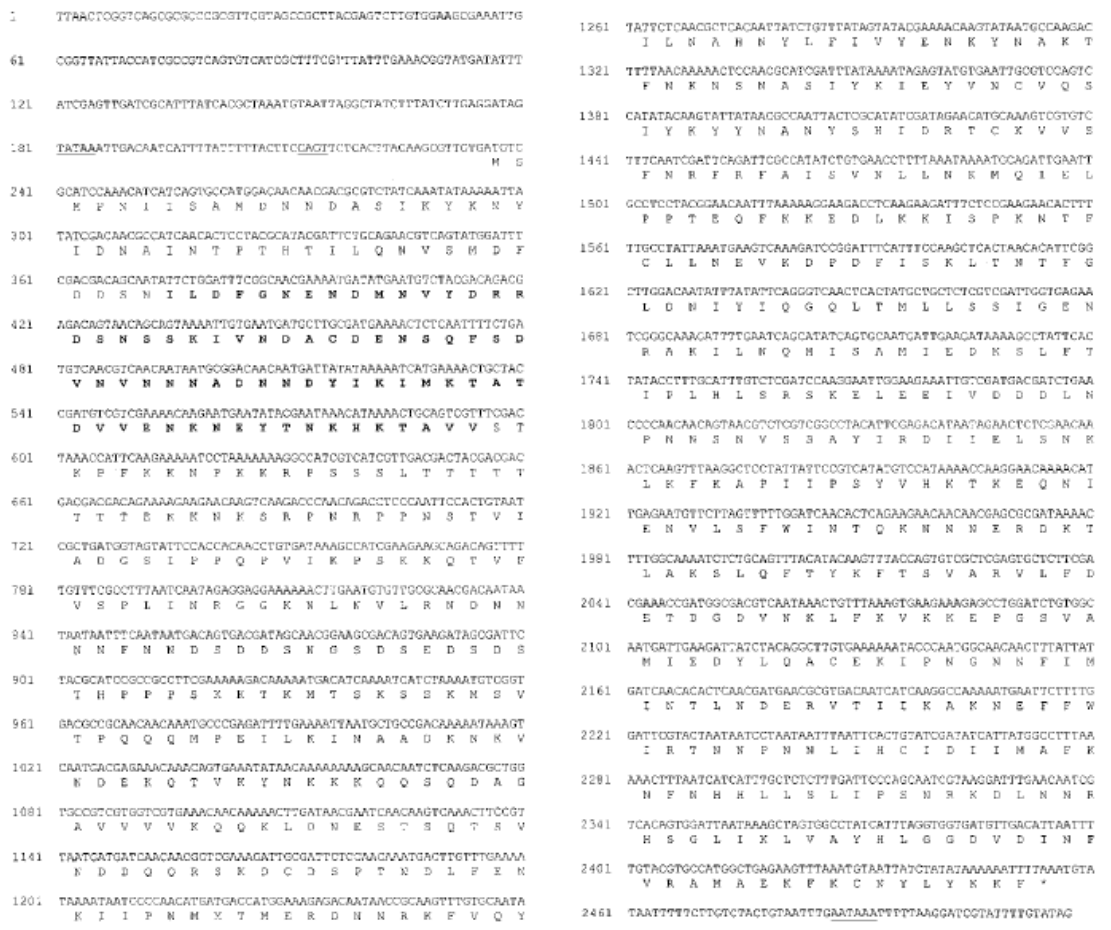

Fig. 2. Nucleotide sequence of the TniSNPV *ie-1* gene and its ORF. The putative ATG is taken as +1. The TATA box, a CAGT motif and a single polyadenylation signal are underlined. The SET-domain-like region of 47–120 amino acids at the N-terminus of the deduced IE-1 protein are in bold.

A comparison of the AcMNPV and OpMNPV IE-1 deduced amino acid sequences revealed that the N-terminal regions were not as conserved as the C-termini regions [10]. In order to see whether this was a general feature of all IE-1 proteins, multiple sequence comparisons were conducted (Table 1). In each case the amino acid sequences of the N-terminal and C-terminal regions were compared with equivalent regions from the other IE-1 proteins. The N-terminal regions consisted of the first third of each protein, while the C-terminal regions consisted of the remaining two thirds. Analysis of the data presented in Table 1 clearly shows that in every case, the C-terminal regions are more conserved than the N-termini. A multiple sequence alignment of the deduced amino acid sequences of IE-1 proteins further indicates that the N-terminal regions are less conserved than the C-terminal regions (Fig. 4).

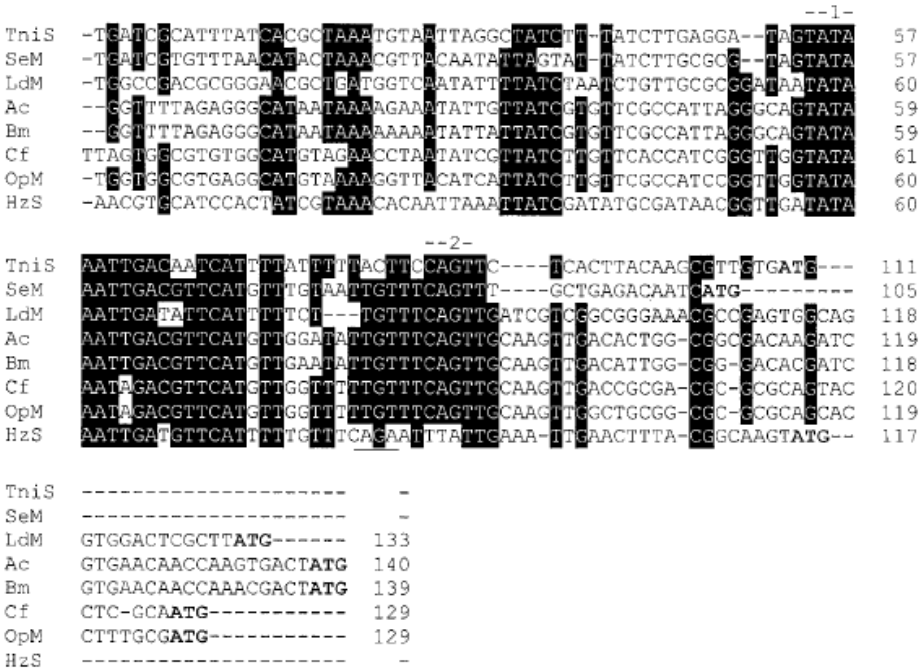

Fig. 3. Multiple sequence alignment of *ie-1* promoter regions. Numbers above the multiple sequence alignments are used to refer to conserved elements. TATA and CAGT elements have been labeled 1 and 2 respectively. A CAGA element of HzSNPV is underlined. Putative ATG translation initiation codons are in bold.

Table 1. Comparison of IE-1 amino acid sequences. Data presented in the tables represents the percentage of identical amino acids shared by the relevant sequences. In each case the N-terminal and C-terminal regions of each IE-1 protein were compared to the equivalent region of every other IE-1 protein. The N-terminal regions consisted of the first third of each protein, while the C-terminal regions consisted of the remaining two-thirds

	BmNPV (%)	CfNPV (%)	OpMNPV (%)	HzSNPV (%)	SeMNPV (%)	LdMNPV (%)	TniSNPV (%)
N-termini							
AcNPV	87	21	23	10	11	9	10
BmNPV		22	23	11	11	8	10
CfNPV			51	8	10	12	9
OpMNPV				6	10	10	10
HzSNPV					11	3	8
SeMNPV						7	11
LdMNPV							7
C-termini							
AcNPV	97	53	53	29	25	26	22
BmNPV		53	52	29	25	26	22
CfNPV			83	29	24	26	22
OpMNPV				28	24	26	22
HzSNPV					33	28	27
SeMNPV						25	27
LdMNPV							22

It has been proposed that the C-terminal regions of IE-1 proteins are involved in DNA binding [10]. Subsequent research suggested that the C-terminus of IE-1 contained aDNA-binding domain [11]. It was further hypothesized that the secondary structure of the C-terminal region may fold into a helix-loop-helix-like DNA binding motif [11]. A multiple sequence alignment was conducted to identify conserved regions within the deduced amino acid sequences of IE-1 proteins (Fig. 4). Six regions were identified in the C-terminal two thirds of the IE-1 proteins. Each region extends over several amino acids with most residues completely conserved. In an attempt to establish the role of these conserved regions the TniSNPV IE-1 protein sequence was compared with entries in the

SwissProt database. However, a significant match to previously characterized domains was not obtained. Therefore the functions of the conserved regions remains to be established.

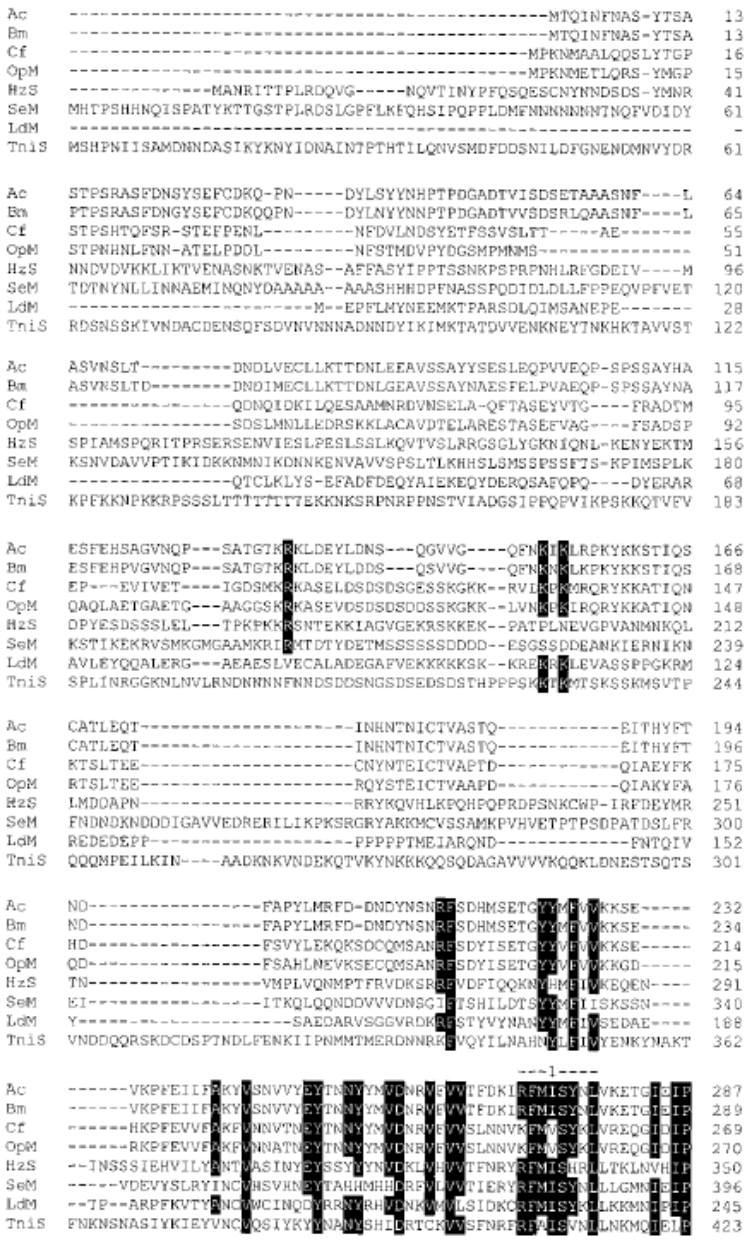

Fig. 4. Alignment of amino acid sequences deduced from nucleotide sequences of eight baculovirus *ie-1* genes. Numbers above the multiple sequence alignments are used to refer to conserved elements. Alignment order is as follows: Ac, *Autographa califonica* NPV (GenBank P11138); Bm, *Bombyx mori* NPV (GenBank P33245); Cf, *Choristoneura fumiferana* NPV (GenBank L04945); Op, *Orgyia pseudotsugata* MNPV (GenBank P22114); Hz, *Helicoverpa zea* SNPV (GenBank U67264); Se, *Spodoptera exigua* MNPV (GenBank AF169823); Ld, *Lymantria dispar* MNPV (GenBank AF006656);. Tni, *Trichoplusia ni* SNPV (GenBank AF317709).*

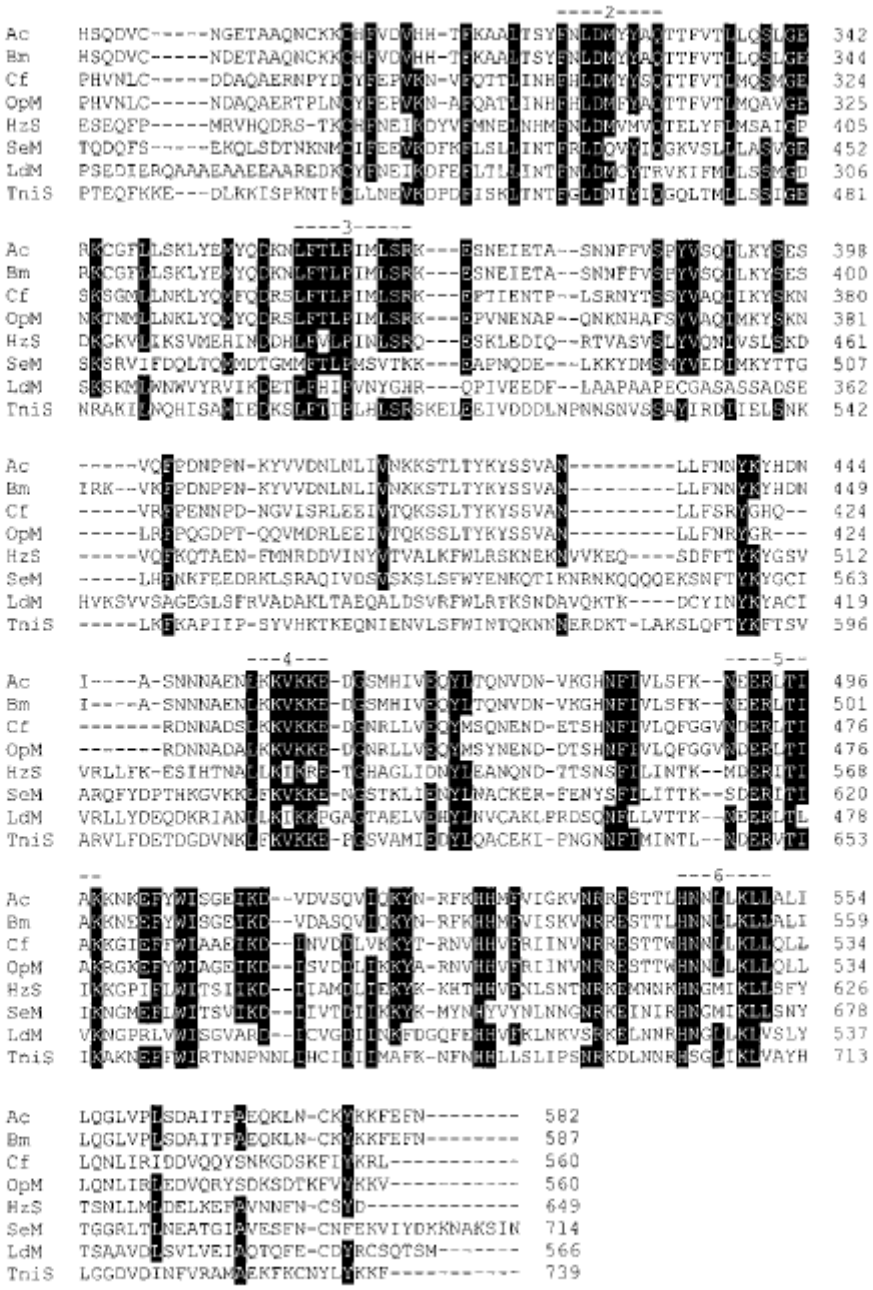

Fig. 4. (Continued)

The role of the N-terminal region of the AcMNPV IE-1 has been experimentally examined [11]. It was demonstrated that a deletion of the first 145 amino acids of the AcMNPV IE-1 protein prevented transcriptional transactivation. However, DNA binding activity was not impaired. While the N-termini of the different IE-1 proteins may share little sequence identity, it was noted that the N-termini for amino acids 1–132 in OpMNPV IE-1 and 1–150 in AcMNPV IE-1 consisted of a high proportion of acidic residues [10]. Also, bordering the acidic region is a cluster of basic amino acids. In this study, a comparison of the charges carried by the amino acids of IE-1 proteins confirmed the N-termini contain similar regions with a high proportion of acidic residues (Fig. 5). A cluster of basic amino acids appears at the boundary of the acidic regions as previously reported [10], but the pattern and position of the clusters are different in each case (Fig. 5).

A comparison was made between the N-terminal region of the TniSNPV IE-1 protein domains within the SwissProt database. This revealed that amino acids 47–120 of the deduced IE-1 protein significantly matched the SET-domain of transcriptional regulators. The SET domain has been found in more than 40 proteins present in organisms ranging from yeasts to mammals. The SET domain is mostly present in chromosomal proteins modulating transcriptional activities and/or chromatin structure [25] and is also probably involved in protein–protein interactions [26].

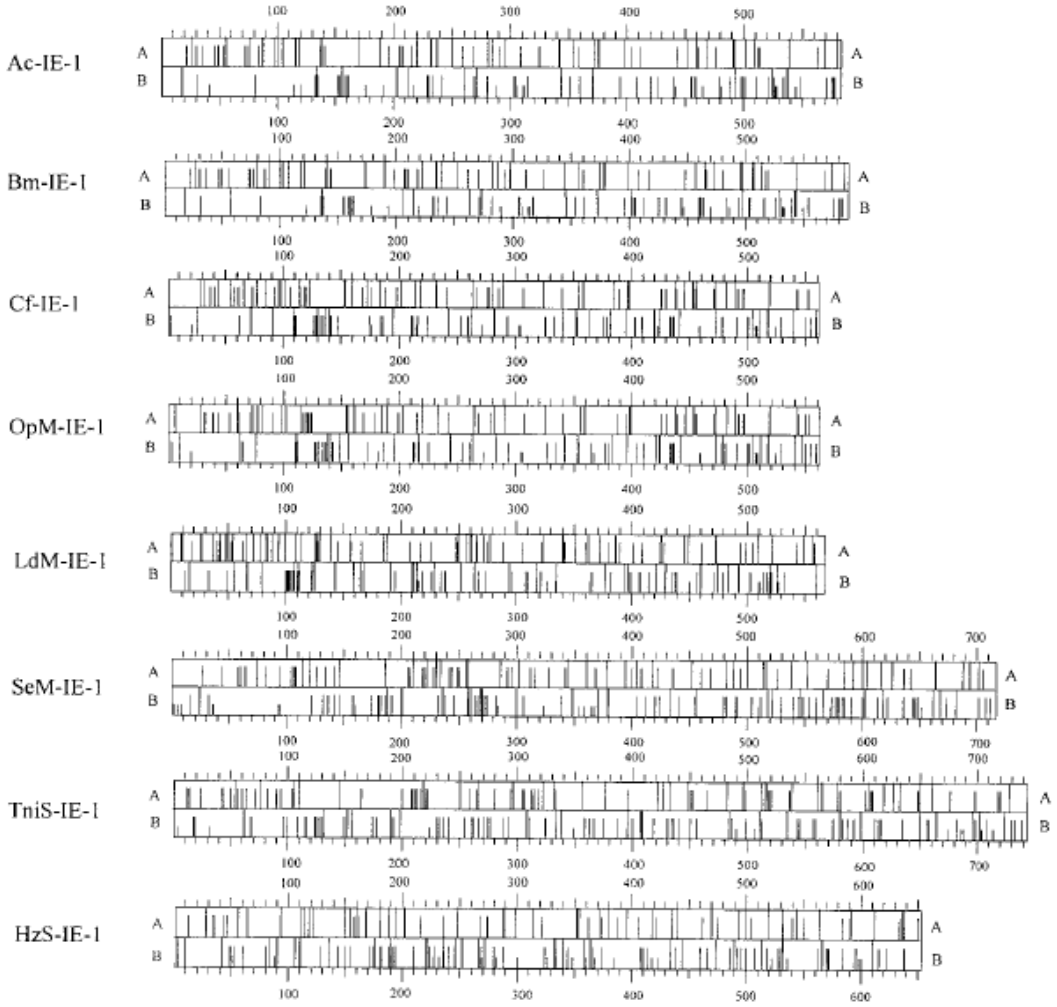

Fig. 5. Acid-base profiles of the eight baculovirus IE-1 proteins. The lane A of each profile refers to the acidic amino acids and lane B to the basic amino acids. Medium bar and full bar represent aspartic acid and glutamic acid in lane A, while small, medium and full bars indicate histidine, lysine and arginine in lane B.

In an attempt to classify TniSNPV phylogeny, a previous study suggested that TniSNPV is phylogenetically related to group II NPVs [17]. This was based on the presence of a unique N-terminal peptide sequence motif MYT(R/P)YS present in the polyhedrin proteins of group II baculoviruses. However, due to the small size of occlusion proteins (245–250 residues) with invariance in more than half [27] and based on comparison of conserved domains only, the phylogenetic estimation is inaccurate. In comparison the NPV *ie-1* may have advantage over polyhydrin for resolving baculovirus phylogeny as it contains a much longer protein sequence. Also, the *ie-1* gene for the Xestia c-nirum GV (XecnGV) provides an appropriate outgroup taxon for rooting the phylogenetic tree.

Table 2. Comparison of the overall amino acid sequences of seven NPV IE-1s with that of TniSNPV IE-1. The highest scores in identity and similarity and the lowest in gap obtained by SeMNPV and HzSNPV are in bold

	AcNPV (%)	BmNPV (%)	CfNPV (%)	OpMNPV (%)	SeMNPV (%)	HzSNPV (%)	LdMNPV (%)
Identity	17	17	16	15	23	21	18
Similarity	31	31	32	31	41	39	33
Gap	21	22	24	24	7	12	27

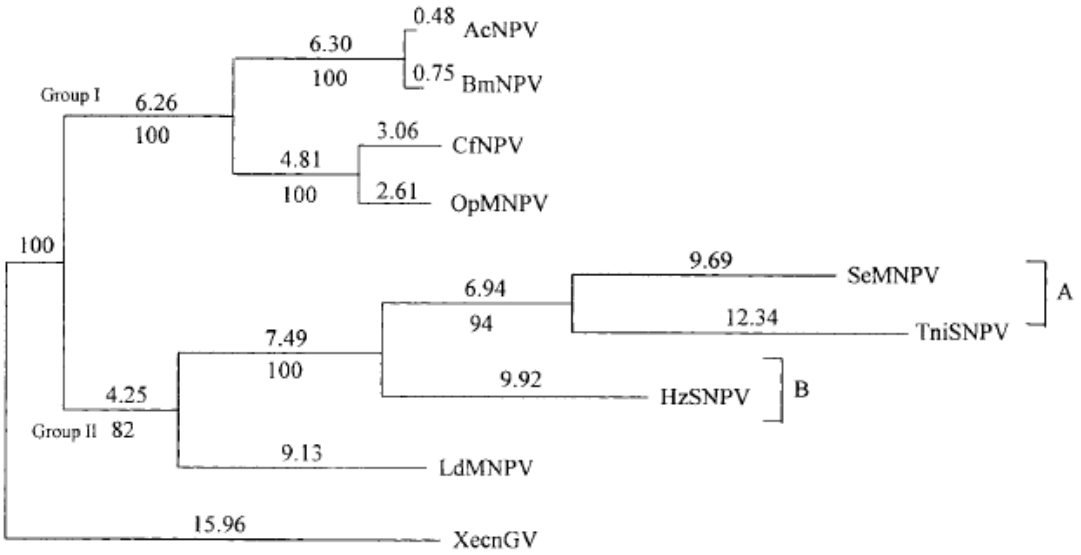

Fig. 6. Phylogenetic tree from the data of nine beculovirus IE-1 proteins was constructed by a heuristic search using PAUP 4.0b4a. The tree length = 2682 steps and consistency index (CI) = 0.9344. Branch lengths above the lines are shown as percentages of the total tree length. Below the lines bootstrap values (100 replicates) from a 50% majority-rule consensus tree are presented. Group I and II are indicated on the tree. The letters, A and B, represent Subgroup II-A and -B respectively. The XecnGV was used as the outgroup.

Comparison of the overall amino acid sequences of the eight NPV *ie-1* genes using Clustalw indicates that the TniSNPV IE-1 shares the highest homologies with group II SeMNPV and HzSNPV IE-1s (Table 2). This sequence alignment data is supported by the estimation of a phylogenetic tree. Figure 6 shows a tree generated using maximum parsimony to estimate the evolutionary relationships based on the protein data set of the *ie-1*s. The input amino acid sequence alignment was created with Clustalw using default parameter settings. Two main branches for the NPVs appeared in this tree, one for Group I taxa and other for Group II taxa. A 82% bootstrapping support was given for inclusion of LdMNPV with the Group II. The inclusion was reported previously [27]. Similarly, SeMNPV and HzSNPV were defined as Subgroups II-A and -B [27] with strong support by their bootstrap values in this tree. The TniSNPV is in a position of subclade II-A with SeMNPV, supported by a 90% bootstrap value. Unfortunately, the tree cannot provide a full picture of Group II due to lack of IE-1 protein data sets for Subclade II-C as defined previously [27] so that it is not conclusive in which Subclade the TniSNPV may belong to. However, the data analyzed by the phylogenetic estimation confirms that TniSNPV is a member of Group II NPVs and also suggested that TniSNPV may be a member of Subgroup II-A.

Acknowledgements
The authors wish to gratefully acknowledge Professor Alan Channing (Department of Zoology, University of the Western Cape) for phylogenetic analysis. This work was funded by the Foundation for Research Development, Pretoria, South Africa and the South African Protea Producers and Exporters Association.

References

1. Kool M., Ahrens C.H., Vlak J.M., and Rohrmann G.F., J Gen Virol *76*, 2103–2118, 1995.
2. Ayres M.D., Howard S.C., Kuzio J., Ferber M.L., and Possee R.D., Virology *202*, 586–605, 1994.
3. Blissard G.W. and Rohrmann G.F., Ann Rev Entomol *35*, 127–155, 1990.
4. Fuchs L.Y.,Woods M.S., andWeaver R.F., J Virol *48*, 641–646, 1983.
5. Kovacs G.R., Guarino L.A., and Summers M.D., J Virol *65*, 5281–5288, 1991.
6. Rodems S.M. and Friesen P.O., J Virol *69*, 5368-5375, 1995.
7. Guarino L.A. and Summers M.D., J Virol *57*, 563–571, 1985.
8. Carson D.D., Summers M.D., and Guarino L.A., Virology *182*, 279–286, 1991.
9. Leisy D.J., Rasmussen C., Owusu E.G., and Rohrmann G.F., J Virol *71*, 5088–5094, 1997.
10. Teilmann D.A. and Stewart S., Virology *180*, 492–508, 1991.
11. Kovacs G.R., Choi J., Guarino L.A., and Summers M.D., J Virol *66*, 7429–7437, 1992.
12. Rodems S.M., Pullen S.S., and Friesen P.D., J Virol *71*, 9270–9277, 1997.
13. Slack J.M. and Blissard G.W., J Virol *71*, 9579–9587, 1997.
14. Murphy F.A., Fauquet C.M., Bioshop D.H.L., Ghabrial S.A., JarvisA.W., Martelli G.P.,MayoM.A., and Summers M.D. (eds). *Virus Taxonomy: The classification and nomenclature of viruses. Six report of the international committee on taxonomy of viruses*. Springer, New York, 1995, pp. 104–113.
15. Zanotto P.M.D.A., Kessing B.D., and Maruniak J.E., J Invertebr. Pathol *62*, 147–164, 1993.
16. Cowan P., Bulach D., Goodge K., Robertson A., and Tribe D.E., J Gen Virol *75*, 3211–3218, 1994.
17. Fielding B.C. and Davison S., Virus Genes *19*, 67–72, 1999.
18. Miller L.K. and Dawes K.P., Appl Environ Microbiol *35*, 411–421, 1977.
19. Henikoff S., Gene *28*, 351, 1984.
20. Altschul S.F., GishW., MillerW., Myers E.W., and Lipman D.J., J Mol Biol *275*, 403–410, 1990.
21. Thompson J.D., Higgins D.G., and Gibson T.J., Nucleic Acids Res *22*, 4673–4680, 1994.
22. Swofford D.L., 2000 PAUP Version 4, Sinauer Associates, Sunderland, Massachusetts.
23. Efron B., CMBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38, Society of Industrial and Applied Mathematics, Philadelphia.
24. Pullen S.S. and Friesen P.D., J Virol *69*, 3575–3583, 1995. 25. Jenuwein T., Laible G., Dorn R., and Reuter G., CMLS Cell Mol Life Sci *54*, 80–93, 1998.
26. Cui X., Vivo I.D., Slany R., Miyamoto A., Firestein R., and Cleary M.L., Nature Genet *18*, 331–337, 1998.
27. Bulach M.D., Kumar C.A., Zaia A., Liang B., and Tribe D.E., J Invertebr Pathol *73*, 59–73, 1999.