Journal of General Virology (2000), **81**, 2313–2325. Printed in Great Britain

Codon usage in nucleopolyhedroviruses

David B. Levin and Beatrixe Whittome

Department of Biology, University of Victoria, Victoria, British Columbia, Canada V8W 3N5

Phylogenetic analyses based on baculovirus polyhedrin nucleotide and amino acid sequences revealed two major nucleopolyhedrovirus (NPV) clades, designated Group I and Group II. Subsequent phylogenetic analyses have revealed three Group II subclades, designated A, B and C. Variations in amino acid frequencies determine the extent of dissimilarity for divergent but structurally and functionally conserved genes and therefore significantly influence the analysis of phylogenetic relationships. Hence, it is important to consider variations in amino acid codon usage. The Genome Hypothesis postulates that genes in any given genome use the same coding pattern with respect to synonymous codons and that genes in phylogenetically related species generally show the same pattern of codon usage. We have examined codon usage in six genes from six NPVs and found that: (1) there is significant variation in codon use by genes within the same virus genome; (2) there is significant variation in the codon usage of homologous genes encoded by different NPVs; (3) there is no correlation between the level of gene expression and codon bias in NPVs; (4) there is no correlation between gene length and codon bias in NPVs; and (5) that while codon use bias appears to be conserved between viruses that are closely related phylogenetically, the patterns of codon usage also appear to be a direct function of the GC-content of the virus-encoded genes.

Introduction

Nucleopolyhedroviruses (NPVs) and granuloviruses (GVs) are members of the family *Baculoviridae*. These viruses infect arthropods and consist of rod-shaped virions that contain large double-stranded, supercoiled DNA genomes ranging in size from 88 to 165 kilobase pairs (kb) (Blissard & Rohrmann, 1990). *Autographa californica* multinucleocapsid NPV (AcMNPV), the most well-characterized baculovirus, has a genome of 134 kb and encodes approximately 150 genes (Ayres *et al.*, 1994). Phylogenetic analyses based on baculovirus polyhedrin nucleotide and amino acid sequences revealed two major NPV clades, designated Group I and Group II (Zanotto *et al.*, 1993).

These analyses suggested that lepidopteran NPVs evolved from a common ancestral virus and that the GVs diverged prior to the divergence of the Group I and Group II NPVs. Subsequent phylogenetic analyses based on the polyhedrin gene (Cowan *et al.*, 1994) and the ecdysteroid UDP-glycosyltransferase (*egt*) gene (Hu *et al.*, 1997) generally supported the phylogenetic relationships suggested by Zanotto *et al.* (1993), but also resulted in conflicting placement of some branches, particularly in the Group II NPVs. More recent phylogenetic analyses of baculoviruses, based on both polyhedrin and DNA polymerase genes (Bulach *et al.*, 1999), revealed three Group II subclades, designated A, B and C.

Variations in amino acid frequencies determine the extent of dissimilarity for divergent but structurally and functionally conserved genes and thus significantly influence the outcome of phylogenetic analyses (Halpern & Bruno, 1998). Hence, it is important to consider variation in amino acid codon usage. The genetic code is degenerate. Most of the 20 amino acids are encoded by more than one codon. Synonymous codon use is distinctly non-random in both prokaryotic and eukaryotic genes (Li, 1997). There is a range of minimal to extreme codon use bias in unicellular organisms (*E. coli* and yeast; Sharp & Li, 1987) and in *Drosophila* species (Shields *et al.*, 1988; Powell & Moriyama, 1997).

The Genome Hypothesis postulates that genes in any given genome use the same coding pattern with respect to synonymous codons. Genes in an organism or in related species generally show the same pattern of codon usage (Li, 1997). In the case of unicellular organisms, the degree of codon use bias is highly correlated with its level of expression in the cell (Gouy & Gautier, 1982; Ikemura, 1985). Highly expressed *E. coli* and yeast genes show a tendency to use a subset of 25

Author for correspondence: David Levin.
Fax +1 250 472 4075. e-mail dlevin@uvic.ca

0001-6972

Table 1. Genes examined, number of amino acids, GenBank accession numbers and references

Virus/gene	No. of amino acids	GenBank accession no.	Reference
AcMNPV			
egt	506	L22858	Ayres *et al.* (1994)
fgf	181		
pk-1	272		
p10	94		
p74	645		
polh	245		
BmNPV			
egt	506	L33180	GenBank submission
fgf	182		
pk-1	275		
p10	78		
p74	645		
polh	245		
OpMNPV			
egt	489	U75930	Ahrens *et al.* (1997)
fgf	205		
pk-1	274		
p10	92		
p74	644		
polh	245		
SpliNPV			
egt	515	AJ003131	GenBank submission
fgf	243	AJ003131	GenBank submission
pk-1	278	X99711	Faktor *et al.* (1997*a*)
p10	104	X99377	Faktor *et al.* (1997*b*)
p74	657	X99376	GenBank submission
polh	249	D01017	Croizier & Croizier (1994)
SpltNPV			
egt	509	X99073	GenBank submission
fgf	239	X99073	GenBank submission
pk-1	254	AF039272	GenBank submission
p10	105	AF037263	GenBank submission
p74	657	AJ011858	GenBank submission
polh	249	AF037262	GenBank submission
LdMNPV			
egt	560	AF081810	Kuzio *et al.* (1999)
fgf	285		
pk-1	274		
p10	77		
p74	672		
polh	245		

and 22 preferred codons, respectively, whereas genes with low levels of expression use codons more randomly (Bennetzen & Hall, 1982).

Codon bias in *Drosophila* genes is consistent over long periods of evolution of *Drosophila* species. With the exception of aspartic acid, all amino acids contribute equally to the codon use bias of a gene. Some *Drosophila* genes, however, do display variant codon bias across species. G and C bases are favoured at synonymous sites in biased genes. Smaller genes tend to have more codon bias than longer genes. Highly and/or rapidly expressed genes tend to be highly biased in their codon usage. The preferred codons in highly biased genes optimally bind the most abundant isoaccepting tRNAs. Thus, codon bias in *Drosophila* genes appears to be driven by translational efficiency, as it is in *E. coli* and yeast (Powell & Moriyama, 1997).

We have examined codon usage in six genes (*egt*, *fgf*, *pk-1*, *p10*, *p74* and *polh*) from six multinucleocapsid nucleopolyhedroviruses [AcMNPV, *Bombyx mori* (Bm)NPV, *Orgyia pseudotsugata* (Op)MNPV, *Spodoptera littoralis* (Spli)MNPV, *Spodoptera litura* (Splt)MNPV and *Lymantria dispar* (Ld)MNPV] to determine if the trends observed in unicellular organisms and in *Drosophila* are conserved within this group of viruses. We have found that: (1) there is significant variation in codon use by genes within the same virus genome; (2) there is significant variation in the codon usage of homologous genes encoded by different NPVs; (3) there is no correlation between the level of gene expression and codon bias in NPVs; (4) there is no correlation between gene length and codon bias in NPVs; and (5) that while codon use bias appears to be conserved between viruses that are closely related phylogenetically, the patterns of codon usage also appear to be a direct function of the GC-content of the virus-encoded genes.

Methods

■ **Nucleotide and amino acid sequences.** The nucleotide and amino acid sequences of six genes (*egt*, *fgf*, *pk-1*, *p10*, *p74* and *polh*) from AcMNPV, BmNPV, OpMNPV, SpliMNPV, SpltMNPV and LdMNPV were downloaded from GenBank. These six genes were selected on the following basis. First, while a large number of genes are available for a few NPVs (AcMNPV, BmNPV, OpMNPV and LdMNPV), we wished to compare codon usage in genes from the largest number of NPVs possible. We selected the six genes listed because they were available from six different viruses. Second, we wished to compare codon usage in genes from NPVs that have different phylogenetic relationships. While many genes from AcMNPV, BmNPV and OpMNPV are available for comparison, these viruses are more closely related to each other phylogenetically (Group I NPVs) than they are to SpliMNPV, SpltMNPV and LdMNPV, which are Group II NPVs. Finally, we wished to compare codon usage in genes that are expressed at different times [early (*egt*, *fgf*), late (*p74*, *pk-1*) and very late (*p10* and *polh*)], and at different levels [low to moderate (*egt*, *p74*, *fgf*, and *pk-1*) versus very high (*p10* and *polh*)] during the infection process. Analysis of codon usage by NPV *dnapol* genes will be published separately

The genes examined, number of amino acids in each gene, GenBank accession numbers and relevant references are listed in Table 1. The GC-content of each gene within each NPV and the average GC-content of the NPV calculated across the six genes analysed are presented in Table 2. The GC-content of NPV genes was determined from GenBank sequences using the DNAStar program.

■ **Codon use analyses.** The GenBank Codon Count program (GenBank, 1999) and the DNAsis program (Pharmacia LKB) were used to calculate codon usage tables for each NPV nucleotide sequence analysed. The Effective Number of Codons (ENCs) for each NPV gene was calculated within Microsoft Office Excel from the codon tables using the formulas described by Wright (1990).

■ **Analyses of codon usage frequency.** The frequency of occurrence of each codon was summed for each of the six NPV genes (*egt*, *fgf*, *p10*, *p74*, *pk-1* and *polh*), from each of the six NPVs, yielding 61 values. The AcMNPV codons were arranged for lowest to highest frequency of occurrence. The percent frequency of each AcMNPV codon was calculated using the total number of codons from the six AcMNPV genes. Using AcMNPV as a base line, the percent differences in codon frequencies observed in the five other viruses were calculated [(% codon frequency NPV n/% codon frequency AcMNPV $- 1) \times 100$] for each codon. These values were compared in a pair-wise manner, in the same codon order, with the AcMNPV base-line. We then compared the percent GC-rich and AT-rich codon frequency usage above and below the AcMNPV codon usage base-line. The frequencies of codon occurrence (in percent), arranged from most AT-rich (i.e. codons with 3 A and/or T nucleotides) to most GC-rich (i.e. codons with 3 G and/or C nucleotides) were also graphed in pair-wise comparisons for viruses with similar GC-content.

Pair-wise differences in codon usage were analysed for statistical significance by Chi-square cumulative analysis using the Syntax1 Statistical Package for Social Sciences (SPSS) for Windows Viewers software. In this analysis, Chi-square values for 61 degrees of freedom $< 80{\cdot}23$ were not statistically significant.

Results

Comparison of ENC values of six genes from six NPV genomes revealed that there is considerable variation in codon usage between genes within individual viral genomes (Fig. 1A, Table 3). For example, ENC values range from a low of 33·9 (*p10*) to a high of 53·5 (*egt*) within the AcMNPV genome, from 35·0 (*p10*) to 55·5 (*fgf*) within the BmNPV genome, from 41·0 (*polh*) to 57·7 (*egt*) within the SpliMNPV genome, and from 27·2 (*p10*) to 43·4 (*pk-1*) within the LdMNPV genome. The ENC values of genes within the OpMNPV genome, ranging from 36·3 (*polh*) to 42·1 (*p10*), are much less variable than those of genes within all other genomes examined.

Comparison of ENC values of the six genes from the six NPV genomes also revealed that there is considerable variation in codon usage by the same gene in different viral genomes (Fig. 1B). ENC values for the *p10* gene, for example, range from a low of 27·2 in LdMNPV to 56·6 in SpliMNPV. The *fgf* gene displays low ENC values in the LdMNPV and SpltMNPV genomes (34·0 and 34·4, respectively) and relatively high ENC values in the SpliMNPV, AcMNPV and BmNPV genomes (46·0, 53·0 and 55·5, respectively). Genes within LdMNPV and OpMNPV have consistently lower ENC values compared to the same genes within other NPVs.

In unicellular organisms and in *Drosophila*, codon use bias is highly correlated with the level of expression in the cell and/or with the length of the gene (Gouy & Gautier, 1982; Bennetzen & Hall, 1982; Ikemura, 1985; Powell & Moriyama, 1997). The *polh* and *p10* genes are the most highly expressed genes in NPVs. Codon usage in *polh* appears biased in OpMNPV and LdMNPV (ENC values of 36·3 and 36·2, respectively; Table 3), and more random in AcMNPV and BmNPV (ENC values of 50·1 and 54·6, respectively; Table 3). Codon usage by *p10* displays the same variability and appears highly biased in LdMNPV, AcMNPV and BmNPV, and more random in SpliMNPV (ENC values of 27·2, 33·9, 35·0 and 56·6, respectively; Table 3). Thus, highly expressed genes such as *polh* and *p10* do not display consistent codon use bias. Comparison of the ENC values displayed by a very short gene

Table 2. GC-content of each gene examined and average GC-content of each NPV

Gene	AcMNPV	BmNPV	SpliMNPV	SpltMNPV	OpMNPV	LdMNPV
egt	45·0	45·0	49·3	47·4	61·1	65·1
fgf	45·6	44·3	49·5	50·8	56·3	67·3
pk-1	37·0	40·7	43·3	41·7	57·7	54·4
p10	43·9	44·1	48·0	48·7	59·1	60·7
P74	47·3	45·8	48·4	50·4	57·6	63·8
polh	49·9	47·3	50·9	50·8	53·3	57·3
Av. %GC …	44·8	44·5	48·2	48·3	57·5	61·4

(*p10*; Table 1) with those of long genes (*egt* and *p74*; Table 1) also failed to reveal a consistent pattern of codon bias (Table 3).

To evaluate the degree of variation in codon usage between the different NPV genomes, we plotted the percent frequency of all codons from the six genes studied, from the lowest to highest occurrence within the AcMNPV genome (Fig. 2). We then superimposed the percent frequencies of each codon for the six genes, in the same order as they occur in the AcMNPV genome, from BmNPV (Fig. 2A), OpMNPV (Fig. 2B), SpliMNPV (Fig. 2C), SpltMNPV (Fig. 2D) and LdMNPV (Fig. 2E). These codon percent frequency comparison plots revealed some very interesting patterns of codon usage in the different NPVs. Fig. 2(A) reveals a relatively low degree of variation in codon usage between AcMNPV (base line) and BmNPV genes. The observed variation, however, was not statistically significant by Chi-square cumulative analysis (Chi-square value = 33·58). Fig. 2(B) reveals a large degree of variation in codon usage between AcMNPV (base-line) and OpMNPV genes, and this variation was statistically significant (Chi-square value = 336·49). Fig. 2(C, D) reveals a moderate degree of variation in codon usage between AcMNPV (base-line) genes and those of SpliMNPV and SpltMNPV genes, respectively. Variations in codon usage between AcMNPV and SpliMNPV, and between AcMNPV and SpltMNPV, were statistically significant (Chi-square value = 159·46 and 107·64, respectively). Fig. 2(E) reveals the variation in codon usage between AcMNPV (base line) and LdMNPV genes. LdMNPV displays the largest variance of codon usage compared with AcMNPV. Chi-square analysis revealed that this variation was also statistically significant (Chi-square value = 659·89).

We then examined the distribution of the variation in codon usage in each NPV compared with that of AcMNPV. Fig. 3 displays the percentage of GC-rich or AT-rich codons used, more frequently or less frequently, by each NPV compared with AcMNPV (base-line). Approximately 47% (46·7%) of the codons within the six genes from BmNPV are used more frequently than those in the AcMNPV genome, while 53·3% of the codons are used less frequently, and of these there is an approximately equal distribution of GC-rich and AT-rich codons. In OpMNPV, 37·7% of the codons are used more frequently, while 62·3% are used less frequently. Of those codons used more frequently, a very large proportion (87·0%) are high GC-content codons. Of those that are used less frequently, 31·6% are high GC-content codons, while 68·4% are high AT-content codons. In SpliMNPV, 49·2% of the codons are used more frequently than those of AcMNPV, while 50·8% of the codons are used less frequently. Of those codons that are used more frequently, 66·7% are high GC-content codons. Of those that are used less frequently 67·7% are high AT-content codons. In SpltMNPV, 54·1% of the codons are used more frequently than those of AcMNPV, while 45·9% of the codons are used less frequently. The majority, 63·6%, of the codons used more frequently are high GC-content codons. Of those that are used less frequently, only 39·3% are high GC-content codons, while 60·7% are high AT-content codons. Finally, in LdMNPV, 45·9% of the codons are used more frequently than those of AcMNPV, while 54·1% of the codons are used less frequently. The vast majority (89·3%) of the codons used more frequently are high GC-content codons. Of those that are used less frequently, only 21·2% are high GC-content codons, while 78·8% are high AT-content codons.

The results of these analyses suggest that codon use variation is a function of GC-content. We tested this hypothesis by plotting pair-wise comparisons of the percent frequency of codons used from NPV genomes of similar GC-content. The percent frequency of all codons from each of the six genes analysed were ordered from the lowest to the highest GC-content (Fig. 4). Fig. 4(A) shows the pair-wise comparison of AcMNPV and BmNPV (average GC-contents of 44·8 and 44·5%, respectively), Fig. 4(B) shows the pair-wise comparison of SpliMNPV and SpltMNPV (average GC-contents of 48·2 and 48·3%, respectively), while Fig. 4(C) shows the pair-wise comparison of OpMNPV and LdMNPV (average GC-contents of 57·5 and 61·4%, respectively). These pair-wise comparisons reveal very close concordance in codon usage between NPVs of similar GC-content, although there is greater variance in codon usage between OpMNPV and LdMNPV. Chi-square cumulative analysis of these pair-wise comparisons revealed that the variations in codon usage between AcMNPV and

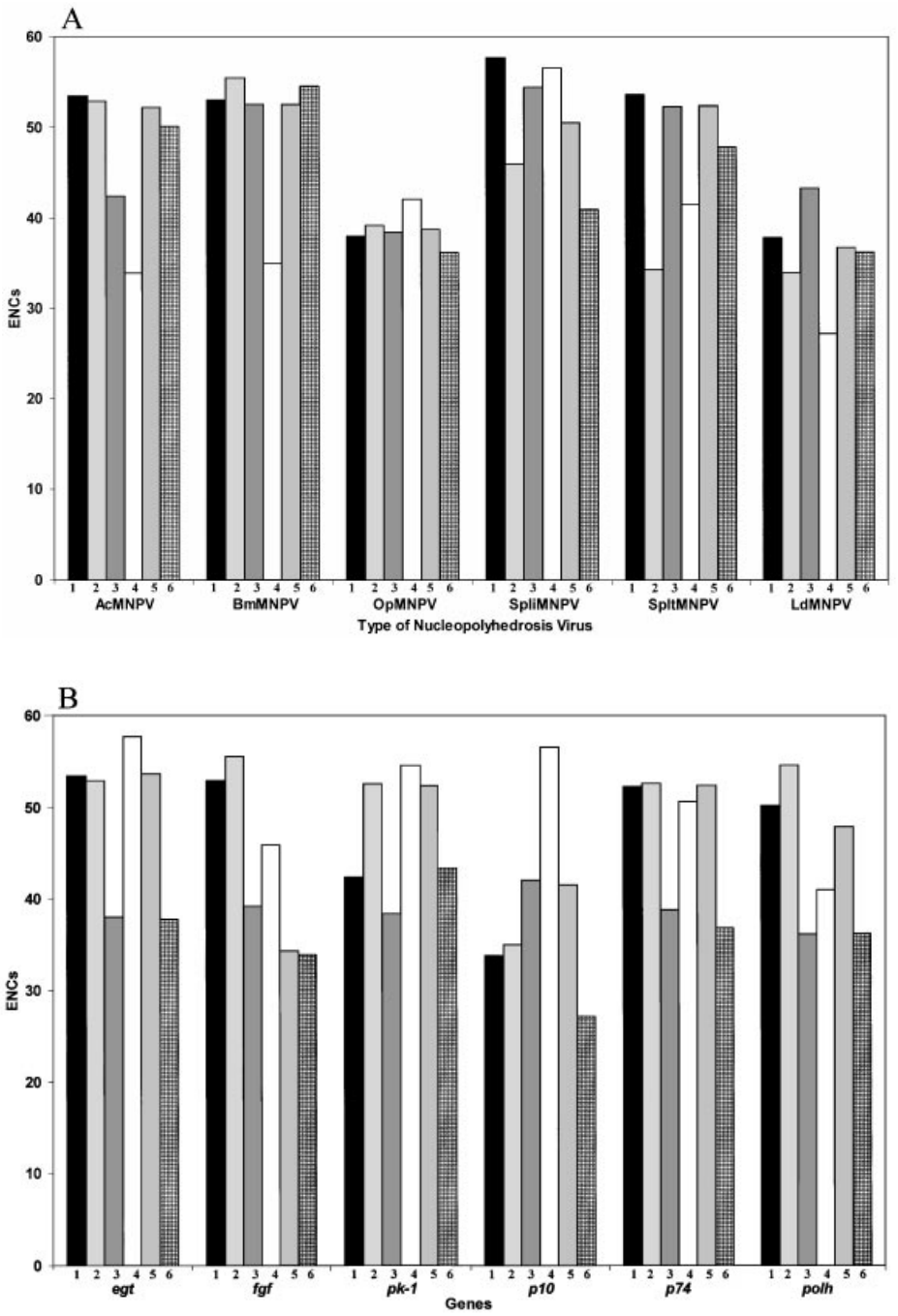

Fig. 1. Comparison of ENC values of six NPV genes within and between NPV genomes. (A) ENC values calculated for the NPV genes were compared within each NPV genome: (1) *egt*; (2) *fgf*; (3) *pk-1*; (4) *p10*; (5) *p74*; and (6) *polh*. (B) ENC values calculated for the six NPV genes (*egt*, *fgf*, *pk-1*, *p10*, *p74* and *polh*) compared between NPVs: (1) AcMNPV; (2) BmMNPV; (3) OpMNPV; (4) SpliMNPV; (5) SpltMNPV; and (6) LdMNPV.

BmNPV and between SpliMNPV and SpltMNPV were not significantly different (Chi-square values of 33·58 and 36·55, respectively), while the variation in codon usage between OpMNPV and LdMNPV was significantly different (Chi-square value = 241·55).

Discussion

Based on previous phylogenetic analyses (Zanotto *et al.*, 1993; Hu *et al.*, 1997; Bulach *et al.*, 1999), AcMNPV, BmNPV and OpMNPV assort as Group I NPVs, while SpliMNPV,

Table 3. Effective number of codons (ENC) values calculated for six genes from AcMNPV, BmNPV, SpliMNPV, SpltMNPV, OpMNPV and LdMNPV

Gene	AcMNPV	BmNPV	SpliMNPV	SpltMNPV	OpMNPV	LdMNPV
egt	53·5	53·0	57·7	53·7	38·0	37·9
fgf	52·9	55·5	46·0	34·3	39·2	34·0
pk-1	42·4	52·6	54·5	52·4	38·4	43·3
p10	33·9	35·0	56·6	41·6	42·1	27·2
p74	52·2	52·6	50·6	52·4	38·8	36·8
polh	50·1	54·6	41·0	47·8	36·3	36·2
Av. ENC …	47·5	50·6	51·1	47·0	38·8	35·9

SpltMNPV and LdMNPV assort as Group II NPVs. The entire nucleotide sequences of the AcMNPV, BmNPV, OpMNPV and LdMNPV genomes have been determined (Ayres *et al*., 1994; GenBank accession no. L33180; Ahrens *et al*., 1997; Kuzio *et al*., 1999; respectively). Comparative genomic analyses have revealed that AcMNPV and BmNPV are very closely related NPVs with greater than 97% overall nucleotide sequence identity. We found that the percentage of GC-nucleotides of each of the six genes examined in this study were very similar for AcMNPV and BmNPV, and that the percentage of GC-nucleotides for these two NPVs, averaged over the six genes, was extremely close (AcMNPV, 44·8%; BmNPV, 44·5%). The similarities in GC-content of AcMNPV and BmNPV genes most probably reflect the close evolutionary relationship between these NPVs. While OpMNPV is a Group I NPV, it is more distantly related to AcMNPV and BmNPV, and has a greater percentage of GC-nucleotides, averaged over the six genes examined (57·5%).

Analysis of a partial sequence of the SpltMNPV DNA polymerase gene placed this virus in the Group II-B subclade (Bulach *et al*., 1999) along with *Helicoverpa armigera* mutlinucleocapsid NPV (HearMNPV) and *H. zea* single nucleocapsid NPV (HzSNPV). Comparison of the SpliMNPV DNA polymerase gene amino acid sequence with the partial amino acid sequence of SpltNPV DNA polymerase gene (our unpublished data) revealed a very high level of sequence identity (94% over 603 amino acids). Phylogenetic analyses, based on both the nucleotide and amino acid sequences, suggest that SpliMNPV and SpltMNPV form a separate subclade, while the other relationships within the Group II NPV delineated by Bulach *et al*. (1999) remain unchanged (our unpublished data). As with AcMNPV and BmNPV, the close similarities between the average GC-content of the six genes examined from SpliMNPV and SpltMNPV (48·2and 48·3%, respectively) most probably reflects the close evolutionary relationship between these NPVs.

While LdMNPV may be considered a Group II NPV, based on phylogenetic analyses of the DNA polymerase gene and polypeptide sequences (Bulach *et al*., 1999), it displays some significant differences in terms of genome size (161 Kb) and genetic organization compared with other NPVs studies to date (Kuzio *et al*., 1999). The much higher average percentage of GC-nucleotides of the six genes examined (61·4%) reflects the overall higher GC-content of LdMNPV.

ENC is a very simple, but effective way of measuring codon use bias (Wright, 1990). ENC is analogous to the effective number of alleles. It is related to the 'homouzygosity' of codons: the probability that two randomly chosen synonymous codons are identical. ENC ranges from 20, if only one codon is used for each amino acid, to 61, if all codons are used equally. ENC values of 35 or less are considered biased and genes with low ENC values are restricted in the use of synonymous codons compared with genes with high ENC values, which have significantly greater flexibility in the use of synonymous codons (Powell & Moriyama, 1997).

We observed that there is significant variation in codon use of genes within the same virus genome and that there is significant variation in the codon usage of homologous genes encoded by different NPVs. Despite the relatively similar average percent GC-content between AcMNPV, BmNPV, SpliMNPV and SpltMNPV, the *p10* genes of AcMNPV and BmNPV have ENC values indicative of codon use bias (33·9 and 35·0, respectively), while the *p10* genes of SpliMNPV and SpltMNPV have ENC values that are indicative of more random codon usage (56·6 and 41·6%, respectively). Moreover, while the average percent GC-contents of OpMNPV and LdMNPV are more similar to each other, and much higher than those of AcMNPV, BmNPV, SpliMNPV and SpltMNPV, the *p10* gene of OpMNPV has an ENC value indicative of random codon usage, while LdMNPV *p10* has an ENC value indicative of extreme bias (42·1 vs 27·2, respectively). Similar patterns were observed for each of the other genes observed in each of the six NPVs. Thus, based on ENC values, codon usage patterns do not appear to be consistent, neither within a particular NPV genome, nor for a particular gene between viral genomes. Nor did we find a correlation between codon use bias and the level of gene expression and/or gene length. Highly expressed genes such as *polh* and *p10* do not display

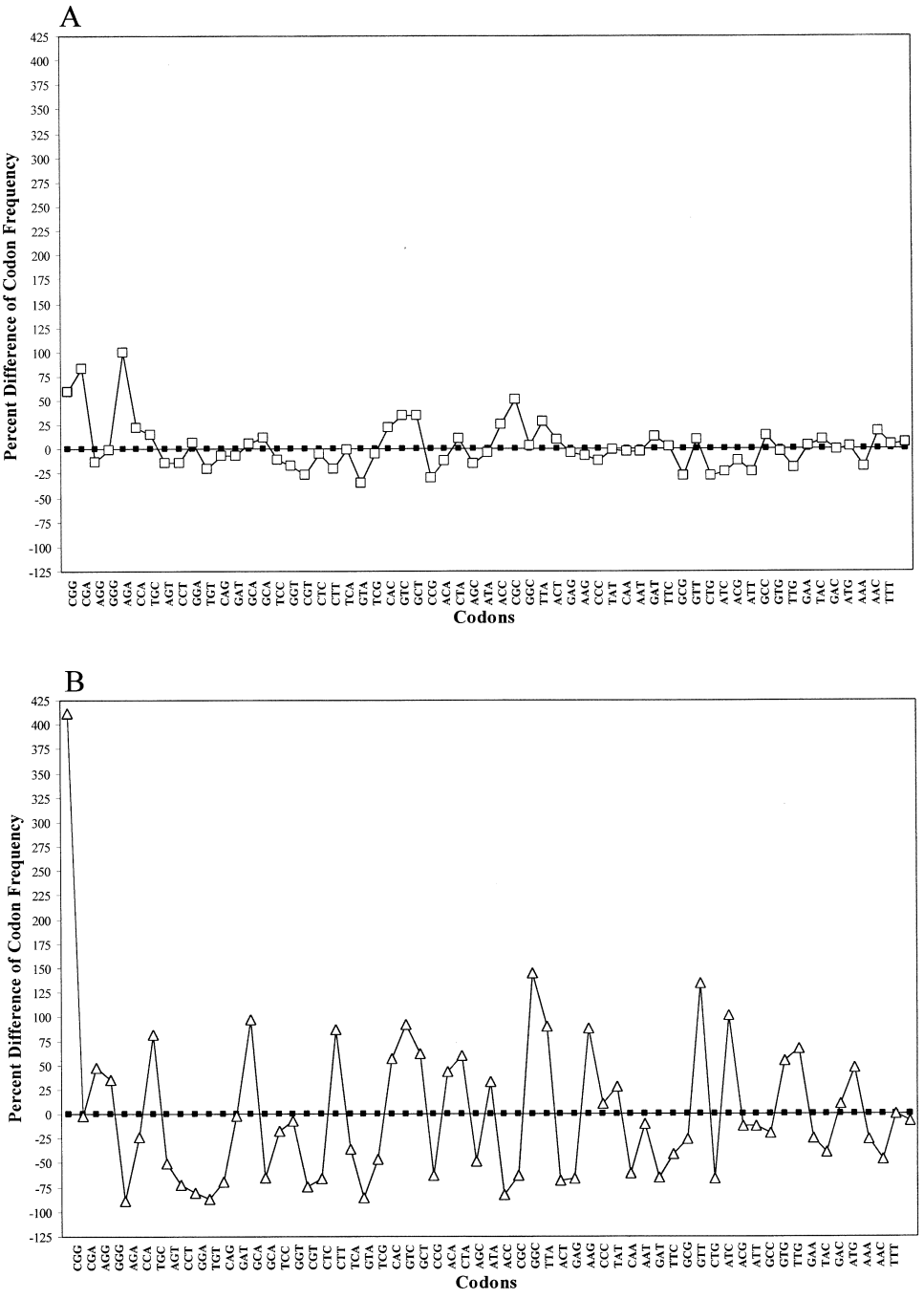

Fig. 2(A, B). For legend see page 2321.

consistent codon use bias. Comparisons of ENC values displayed by a very short gene (*p10*) with those of long genes (*egt* and *p74*) also failed to reveal a consistent pattern of codon bias. These observations are consistent with a previous analysis of codon usage of AcMNPV genes (Ranjan & Hasnain, 1995), in which it was found that certain genes displayed significant variation from the overall AcMNPV codon use pattern. While the *p10* and *polh* genes displayed the greatest variation from the overall codon usage profile, genes from both the early (*ie-2*) and late (*39k*, *sod*) gene classes also displayed variations in codon usage.

Taking the GC-content of each gene within each virus into account, however, the pair-wise comparisons of codon usage in AcMNPV with codon usage for each of the other NPVs revealed that the greater the GC-content difference between AcMNPV and the other NPVs examined, the greater the

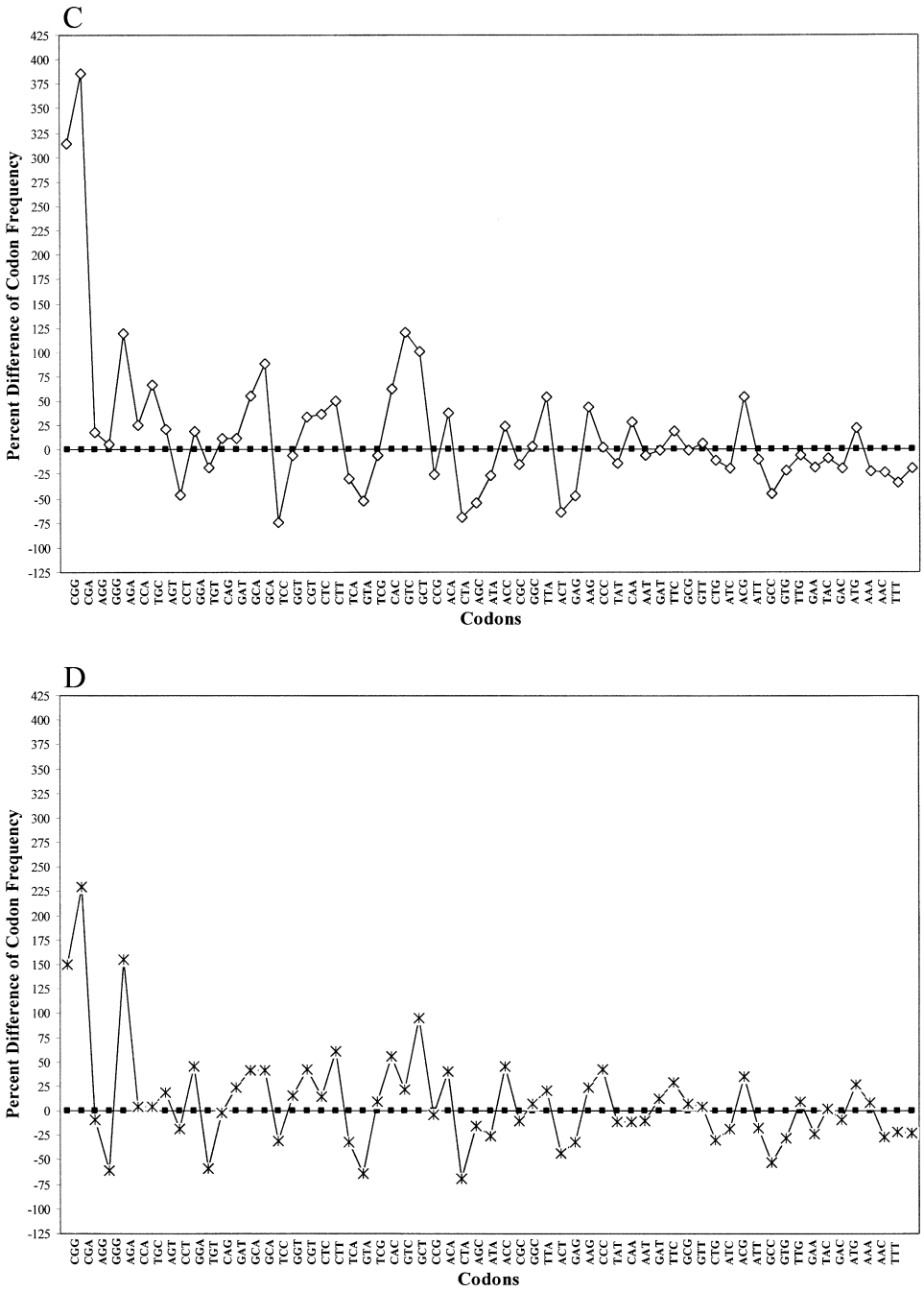

Fig. 2(C, D). For legend see facing page.

variance in codon usage. The variance in codon usage between AcMNPV and BmNPV, which have very similar GC-contents (44·8 and 44·5%, respectively) averaged across the six genes used in this analysis (a difference in average GC-content of only 0·3%), was not statistically different (Chi-square value = 33·58). In contrast, variance in codon usage between AcMNPV and LdMNPV, which differ greatly in GC-content (44·8 vs 61·4%, a difference of 16·6%), was highly significant statistically (Chi-square value = 659·89). LdMNPV codons that were used more frequently tended to be GC-rich codons, while LdMNPV codons that were used less frequently tend to be AT-rich codons, a phenomenon also observed in unicellular organisms and *Drosophila*, where GC-nucleotides are favoured at synonymous sites (Powell & Moriyama, 1997).

Pair-wise comparisons of NPVs of similar GC-content (AcMNPV vs BmNPV; SpliMNPV vs SpltMNPV; OpMNPV vs LdMNPV) also revealed that NPVs with very similar GC-content had highly similar patterns of codon usage. Two pairs

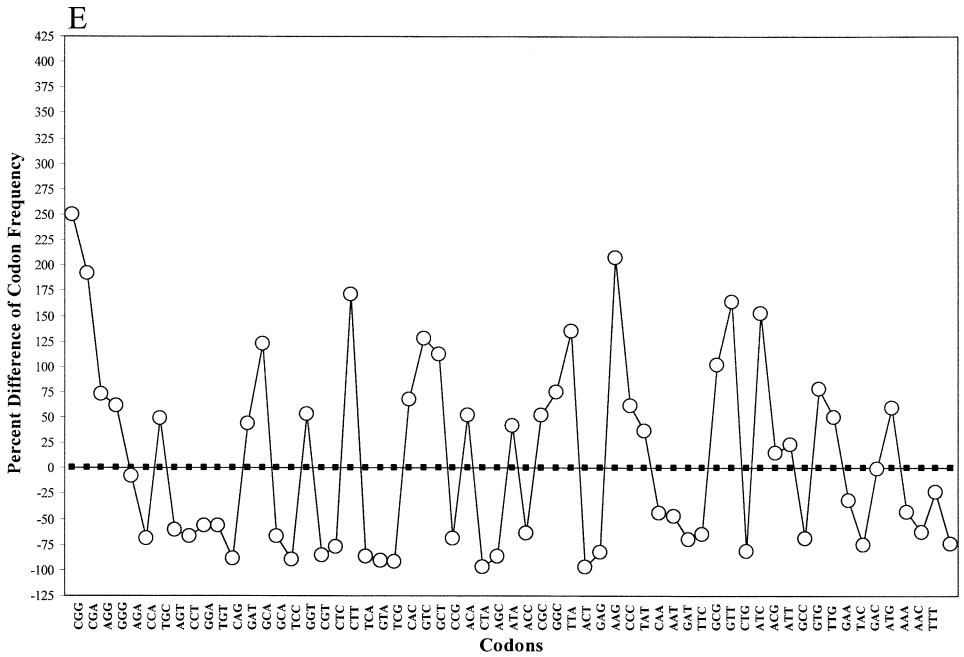

Fig. 2. Pair-wise comparison of the percent frequency difference in codon usage from six NPV genes with codons ordered from lowest to highest occurrence within the AcMNPV genome. The percent frequency of all codons from the six genes (*egt, fgf*, *pk-1, p10, p74* and *polh*) were plotted from the lowest to the highest occurrence within the AcMNPV genome (heavy base-line with solid squares). The percent frequency differences of each codon from the six genes from each NPV were then superimposed, in the same order, in pair-wise comparisons with the AcMNPV codon use pattern: (A) AcMNPV vs BmNPV (light line with open squares); (B) AcMNPV vs OpMNPV (light line with open triangles); (C) AcMNPV vs SpliMNPV (light line with open diamonds); (D) AcMNPV vs SpltMNPV (light line with Xs); (E) AcMNPV vs LdMNPV (light line with open circles).

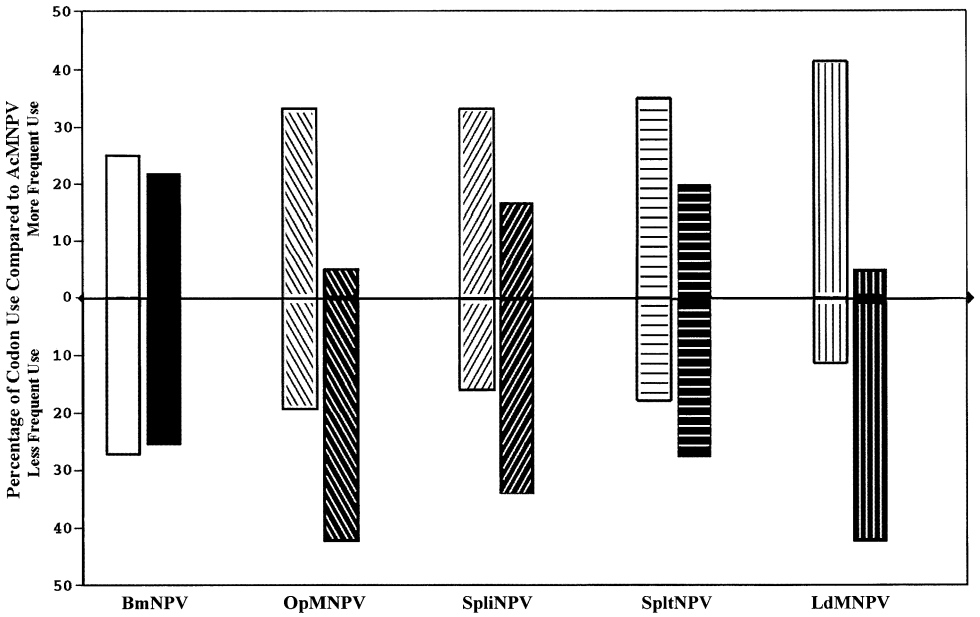

Fig. 3. Percentage of GC-rich or AT-rich codons, used more frequently or less frequently, by each NPV compared with AcMNPV. The percentage of GC-rich (light columns) and AT-rich (dark columns) codons used more (above the AcMNPV base-line) and less frequently (below the AcMNPV base-line) was plotted for each NPV.

of closely related NPVs (AcMNPV and BmNPV; SpliMNPV and SpltMNPV) displayed extremely high concordance in codon use patterns, and the variation in codon usage between each pair was not significantly different by Chi-square cumulative analysis (Chi-square values of 33·58 for AcMNPV vs BmNPV and of 36·55 for SpliMNPV vs SpltMNPV).

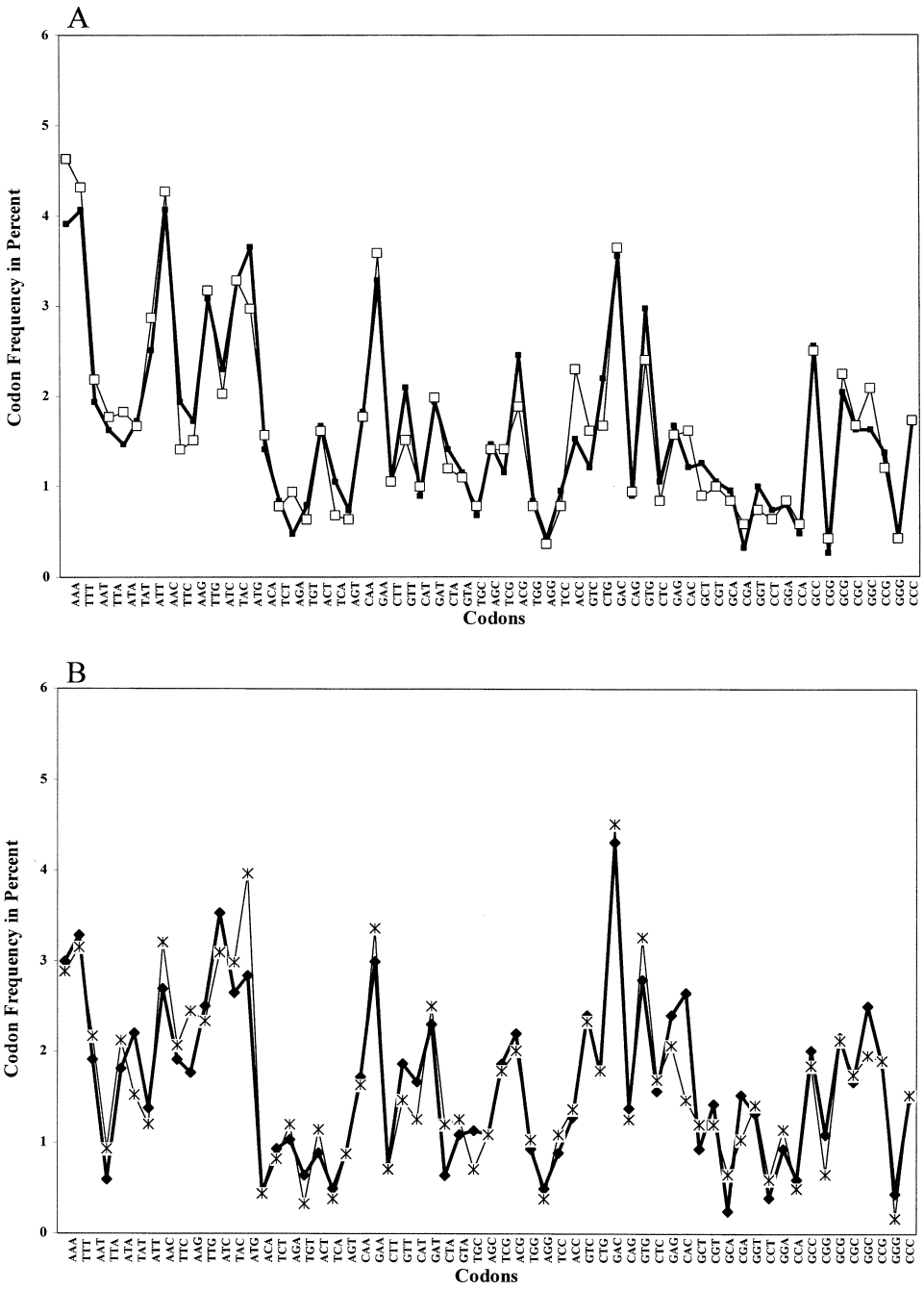

Fig. 4(A, B). For legend see facing page.

The similarities in codon usage between these NPVs, however, may not be attributable to phylogenetic relatedness alone. OpMNPV is phylogenetically more similar to AcMNPV and BmNPV than to LdMNPV. While the variation in codon usage between OpMNPV and LdMNPV was statistically significant (Chi-square value = 241·55), OpMNPV and LdMNPV have GC-contents and patterns of codon usage that are much more similar to each other than does OpMNPV compared with AcMNPV and BmNPV. Similarly, while the variation in codon usages between SpliMNPV (a Group II NPV) and AcMNPV, and between SpliMNPV and BmNPV, are significantly different (Chi-square values of 136·91 and 127·91, respectively), the pattern of codon usage displayed by SpliMNPV is more similar to AcMNPV and BmNPV than it is to LdMNPV. The same is true of SpltMNPV when compared with AcMNPV and BmNPV. Thus, while codon use bias appears to be conserved between viruses that are closely related phylogenetically, the patterns of codon usage also appear to be a direct function of the GC-content of the virus-encoded genes.

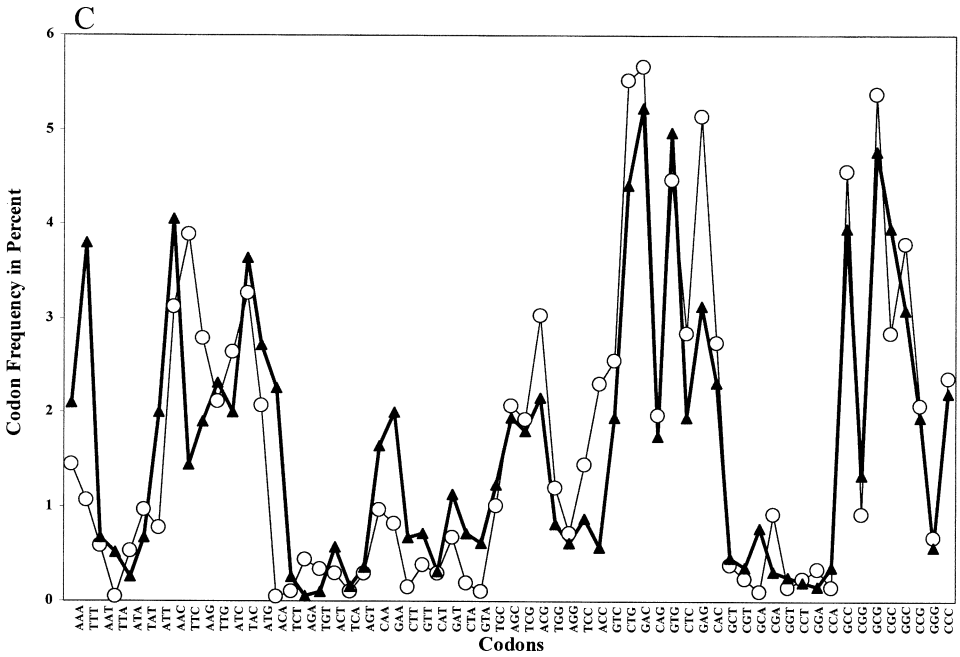

Fig. 4. Pair-wise comparison of the percent frequency difference in codon usage between NPVs with similar GC-content with codons ordered from lowest to highest GC-content. The percent frequency of all codons from the six genes (*egt*, *fgf*, *pk-1*, *p10*, *p74* and *polh*), plotted from the lowest to the highest GC-content, were compared with the percent frequency of codons from each NPV of similar GC-content: (A) AcMNPV (heavy line with solid squares) vs BmNPV (light line with open squares); (B) SpliMNPV (heavy line with solid diamonds) vs SpltMNPV (light line with Xs); (C) OpMNPV (heavy line with solid triangles) vs LdMNPV (light line with open circles).

The overall conclusion from this study is that codon usage within NPV genes does not follow the patterns observed from *E. coli*, *yeast* and *Drosophila*. These organisms, however, are not viruses and codon usage has been analysed in only a few viruses. Human immunodeficiency virus type 1 (HIV-1) and other lentiviruses demonstrate an inverse correlation between the extent of codon bias and the rate of translation of the viral genes. The *pol* genes of lentiviruses display a shift toward AT-rich codons (Bronson & Anderson, 1994). The *gag* genes do not display this bias toward AT-rich codons, despite the fact that expression of the *gag* genes exceeds that of the *pol* genes by a factor of 20, due to infrequent frame-shifting during translation of the *gag–pol* mRNA. It was hypothesized that the aminoacyl-tRNA availability within the host cell restricts the lentivirus preference for AT-rich codons (van Hemert & Berkhout, 1995).

Analyses of bovine papillomavirus (BPV) late gene expression have demonstrated that BPV late genes (capsid proteins L1 and L2) mRNAs are efficiently translated in non-replicating, terminally differentiated keratinocytes, but not in actively dividing undifferentiated cells (Stoler *et al.*, 1992). Zhou *et al.* (1999) hypothesized that BPV late mRNAs may not be efficiently translated in dividing, undifferentiated cells due to a mis-match between codon usage of the viral genes and aminoacyl-tRNA availability in host cell. They suggest that translation of BPV capsid genes may be limited in actively dividing cells due to competition between viral and host mRNAs for rare tRNAs. In terminally differentiated cells, the number of competing host mRNAs would be greatly reduced, permitting increased levels of translation of the viral mRNAs (Zhou *et al.*, 1999).

BPV capsid genes, however, are translated efficiently in insect cells using the AcMNPV baculovirus expression system (BEVS) (Volpers *et al.*, 1994), which is used for high-level synthesis of heterologous proteins (O'Reilly *et al.*, 1992). Efficient translation of BPV capsid gene mRNAs in NPV-infected cells may be due to the lack of competition between host mRNAs and the BPV mRNAs because NPV infections result in a shut-down of host cell mRNA and protein synthesis (Ooi & Miller, 1988; Carstens *et al.*, 1979). Thus, one could argue that the lack of consistency in codon bias displayed by NPV genes, as well as the lack of correlation between codon usage and expression levels of NPV genes, may be due to a lack of competition between host and viral mRNAs for aminoacyl-tRNAs. Selection for translational efficiency in NPV genes would be minimal in NPV-infected cells as aminoacyl-tRNAs would not be a limiting factor. The strong codon bias displayed by some NPV genes, *p10* in LdMNPV for example, may simply reflect the strong GC-bias in these genomes.

Our data suggest that NPVs display a wide variation in codon usage which, when coupled with a lack of competition between host and viral mRNAs for aminoacyl-tRNAs, provides a working hypothesis to explain why the BEVS are so useful for high-level expression of heterologous proteins. Not

all heterologous proteins, however, are successfully expressed at high levels. The beta subunit of human chorionic gonadotropin (*βhCG*) is expressed at very low levels in the AcMNPV BEVS compared with other heterologous proteins. A comparison of the *βhCG* codon usage pattern with the codon use patterns of a heterologous protein that is expressed at high levels [the firefly luciferase (*luc*) gene product], and with the codon usage patterns of numerous AcMNPV genes, revealed that *βhCG* displayed a significantly distinct codon usage that could account for its low level of expression (Ranjan & Husnain, 1994). Moreover, a comparison of the *βhCG* and AcMNPV consensus translation initiation codon sequences revealed major differences that could also influence the low level of *βhCG* expression (Ranjan & Husnain, 1994). Thus, the basis for the differences in the levels of heterologous protein expression in BEVS may be a consequence of several factors, including extreme bias in codon usage and differences in translation initiation codon context.

We would like to thank Dr Tom Reimchen for his advice concerning statistical analysis and Graham Sinclair for his valuable comments and helpful suggestions concerning codon usage in viruses. We also thank Jianhe Huang, from our laboratory, for the SpliMNPV DNA polymerase data and phylogenetic analyses. This work was supported in part by a Canadian Forest Service/NSERC-Industrial Research Grant.

References

Ahrens, C. H., Russell, R. L., Funk, C. J., Evans, J. T., Harwood, S. H. & Rohrmann, G. F. (1997). The sequence of the *Orgyia pseudotsugata* multinucleocapsid nuclear polyhedrosis virus genome. *Virology* **229**, 381–399.

Ayres, M. D., Howard, S. C., Kuzio, J., Lopez-Ferber, M. & Possee, R. D. (1994). The complete DNA sequence of *Autographa californica* nuclear polyhedrosis virus. *Virology* **202**, 586–605.

Bennetzen, J. L. & Hall, B. D. (1982). Codon selection in yeast. *Journal of Biological Chemistry* **257**, 3026–3031.

Blissard, G. W. & Rohrmann, G. F. (1990). Baculovirus diversity and molecular biology. *Annual Review of Entomology* **35**, 127–155.

Bronson, C. B. & Anderson, J. N. (1994). Nucleotide composition as a driving force in the evolution of retroviruses. *Journal of Molecular Evolution* **38**, 506–532.

Bulach, D. M., Kumar, A., Zaia, A., Liang, B. & Tribe, D. E. (1999). Group II nucleopolyhedrovirus subgroups revealed by phylogenetic analyses of polyhedrin and DNA polymerase gene sequences. *Journal of Invertebrate Pathology* **73**, 59–73.

Carstens, E., Tija, S. T. & Doerfler, W. (1979). Infection of *Spodoptera frugiperda* cells with *Autographa californica* nuclear polyhedrosis virus. I. Synthesis of intracellular proteins after infection. *Virology* **99**, 386–398.

Cowan, P., Bulach, D., Goodge, K., Robertson, A. & Tribe, D. E. (1994). Nucleotide sequence of the polyhedrin gene of *Helicoverpa zea* single nucleocapsid nuclear polyhedrosis virus: placement of the virus in lepidopteran nuclear polyhedrosis virus group II. *Journal of General Virology* **75**, 3211–3218.

Croizier, L. & Croizier, G. (1994). Nucleotide sequence of the polyhedrin gene of *Spodoptera littoralis* multiple nucleocapsid nuclear polyhedrosis virus. *Biochimica et Biophysica Acta* **1218**, 457–459.

Faktor, O., Toister-Achituv, M. & Nahum, O. (1997*a*). Enhancer element, repetitive sequences and gene organization in an 8-kbp region containing the polyhedrin gene of the *Spodoptera littoralis* nucleopolyhedrovirus. *Archives of Virology* **142**, 1–15.

Faktor, O., Toister-Achituv, M., Nahum, O. & Kamensky, B. (1997*b*). The *p10* gene of *Spodoptera littoralis* nucleopolyhedrovirus: nucleotide sequence, transcriptional analysis and unique gene organization in the *p*10 locus. *Journal of General Virology* **78**, 2119–2128.

GenBank (1999). Codon Count Program 112.0. Codon Usage Database released by GenBank http://www.kazusa.or.jp/codon.

Gouy, M. & Goutier, C. (1982). Codon useage in bacteria: correlation with expressivity. *Nucleic Acids Research* **10**, 7055–7074.

Halpern, A. L. & Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular Biology and Evolution* **15**, 910–917.

Hu, Z. H., Broer, R., Westerlaken, J., Martens, J., Jin, F., Jehle, J. A., Wang, L. M. & Vlak, J. M. (1997). Characterization of the ecdysteroid UDP-glycosyltransferase gene of a single nucleocapsid nucleopolyhedrovirus of *Buzura suppressaria*. *Virus Research* **47**, 91–97.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution* **2**, 13–34.

Kuzio, J., Pearson, M. N., Harwood, S. H., Funk, C. J., Evans, J. T., Slavicek, J. M. & Rohrmann, G. F. (1999). Sequence and analysis of the genome of a baculovirus pathogenic for *Lymantria dispar*. *Virology* **253**, 17–34.

Li, W.-H. (1997). *Molecular Evolution*, 2nd edn. Sunderland, MA: Sinauer Associates.

Ooi, B. G. & Miller, L. K. (1988). Regulation of host RNA levels during baculovirus infection. *Virology* **166**, 515–523.

O'Reilly, D. R., Miller, L. K. & Lucknow, V. A. (1992). *Baculovirus Expression Vectors: A Laboratory Manual*. New York: W. H. Freeman.

Powell, J. R. & Moriyama, E. N. (1997). Evolution of codon bias in *Drosophila*. *Proceedings of the National Academy of Sciences, USA* **94**, 7784–7790.

Ranjan, A. & Hasnain, S. E. (1994). Influence of codon usage and translation initiation context in the AcMNPV-based expression system: computer analysis using homologous and heterologous genes. *Virus Genes* **9**, 149–153.

Ranjan, A. & Hasnain, S. E. (1995). Codon usage in the prototype baculovirus: *Autographa californica* nuclear polyhedrosis virus. *Indian Journal of Biochemistry and Biophysics* **32**, 424–428.

Sharp, P. M. & Li, W. H. (1987). An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution* **24**, 28–38.

Shields, D. C., Sharp, P. M., Higgins, D. G. & Wright, F. (1988). Silent sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Molecular Biology and Evolution* **5**, 704–716.

Stoler, M. H., Rhodes, C. R., Whitbeck, A., Wollinsky, S. M., Chow, L. T. & Broker, T. R. (1992). Human papillomavirus type 16 and 18 genes expression in cervical neoplasias. *Human Pathology* **23**, 117–128.

van Hemert, F. J. & Berkhout, B. (1995). The tendency of lentiviral open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. *Journal of Molecular Evolution* **41**, 132–140.

Volpers, C., Schirmacher, P., Streek, E. & Sapp, M. (1994). Assembly of the major and the minor capsid protein of human papillomavirus type 33 into virus-like particles and tubular structures in insect cells. *Virology* **200**, 504–512.

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene* **87**, 23–29.

Zanotto, P. M., Kessing, B. D. & Maruniak, J. E. (1993). Phylogenetic relationships among baculoviruses: evolutionary rates and host associations. *Journal of Invertebrate Pathology* **62**, 47–164.

Zhou, J., Liu, W. J., Peng, S. W., Sun, X. Y. & Fraser, I. (1999). Papillomavirus capsid protein expression depends on a match between codon usage and tRNA availability. *Journal of Virology* **73**, 4972–4982.

Received 8 February 2000; Accepted 10 May 2000