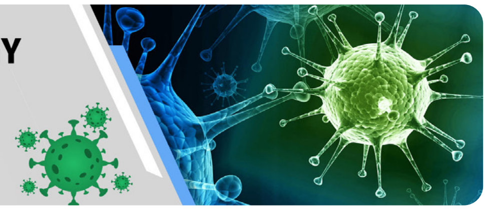
fortune Journals

Research Article

ARCHIVES OF MICROBIOLOGY & IMMUNOLOGY

ISSN: 2572-9365



Effect of Natural Selection on the Codon Usage pattern of alphabaculovirus Genomes

Puttatida Mahapattanakul and Patsarin Rodpothong Wongkamhang*

Abstract

Codon usage is a reflection of evolutionary adaptation to environmental pressure. Codon usage pattern may be unique to species of viruses, genomes of the same species or genes within the same genome. Here, we have analyzed the overall nucleotide composition and the nucleotide composition at the three codon positions in the genomes of 6 alphabaculoviruses. The results suggest that the alphabaculovirus genomes are predominantly under an influence of a natural selection that bias toward A/T. Principle Component Analysis (PCA) based on Relative Synonymous Codon Usage (RSCU) of all Open Reading Frames (ORFs) was employed to investigate the pattern of the codon usage. The majority of the alphabaculovirus ORFs, except those of AgipMNPV, clusters at the same location in the 2-dimensional PCA plot, indicating similar RSCU indices and supporting our hypothesis that genomes of the related viruses should possess similar codon usage pattern because they share similar mechanisms of virus replication and infection, thus are subjected to the same type of evolutionary pressure. A distinct pattern of the RSCU index in the *p6.9* gene, which is an outlier in the PCA plot, also demonstrates an influence of strong natural selection and a type of selection pressure that reflects its functional conservation in DNA packaging as well as its evolutionary relation to protamine-like gene.

Keywords: Codon usage, Alphabaculoviruses, Mutational bias, Natural selection

Introduction

The baculoviruses (family: *Baculoviridae*) are a group of large double-stranded DNA arthropod-specific viruses. They can be categorized into four genera; *Alphabaculovirus*, *Betabaculovirus*, *Gammabaculovirus* and *Deltabaculovirus*. Baculoviruses can also be classified into two types, nucleopolyhedroviruses (NPVs) and granuloviruses (GVs), based on their occlusion bodies (OBs) produced at the late stages of infection [1]. The OB is an organized structure, composed of polyhedrin, which provides stability to virions embedded within, and is responsible for virus horizontal transmission among their insect hosts [2, 3]. Genera *Alphabaculovirus*, *Gammabaculovirus* and *Deltabaculovirus* consist of NPVs, infecting insects belonging to the orders *Lepidoptera*, *Hymenoptera* and *Diptera* [4], while the genus *Betabaculovirus* consists of GVs and only infects lepidopteran insects. Genomes of baculoviruses range from 80–180 kbp in size, encoding 90-180 genes. There are 38 genes that are conserved across different genera of baculoviruses and have been assigned as "core genes", involving in viral DNA replication and packaging, transcription, architecture and assembly [6-9, 32]. Baculoviruses confer high degree of host specificity and insecticidal

Affiliation:

Mahidol International College, Mahidol University, Salaya, Nakhon-Pathom, Thailand

***Corresponding author:**
Patsarin Rodpothong Wongkamhang, Mahidol International College, Mahidol University, Salaya, Nakhon-Pathom, Thailand.

Citation: Puttatida Mahapattanakul and Patsarin Rodpothong Wongkamhang. Effect of Natural Selection on the Codon Usage pattern of alphabaculovirus Genomes. Archives of Microbiology and Immunology 6 (2022): 274-283.

Received: December 09, 2022
Accepted: December 15, 2022
Published: December 24, 2022

activity, thus various NPVs are being studied and developed as environmental friendly biological pesticides that can be effectively used for pest management in agriculture and forestry [10]. Baculoviruses have also been used extensively in cell expression system in the production of recombinant proteins [11, 12]. Codon usage pattern reflects evolution of genes and genomes. Two evolutionary forces that have been suggested to shape the codon usage are neutral mutation and natural selection (14-16). The neutral mutation model states that codon usage bias arises from a bias in nucleotide composition, which in turn arises from a bias in the rate of point mutation, or a bias in the DNA repair mechanism. For example, point mutations that favor the change from A to G and T to C may give rise to a GC-rich genome. It is deemed as !neutral" because these changes often occur at the third codon position, thus do not affect the amino acid sequence and has no fitness advantage. In contrast, the natural selection model suggests that these synonymous mutations would influence the fitness of an organism, such as accuracy and efficiency of translation, and therefore be promoted or repressed during evolution. Previous studies have shown that both neutral mutation and natural selection exert different degrees of influence on different viral genes and genomes. [13] suggested that mutational pressure rather than natural selection is the main determinant of codon usage in vertebrate-infecting DNA viruses [13]. [14] also suggested that mutation is the most important determinant of the codon bias in human RNA viruses, but also proposed that translational selection many have some influence in shaping codon usage bias [14]. Chen 2013 showed that 27% and 21% of total variation in the codon usage pattern could be attributed to mutational pressure, while 5% and 6% of total variation could be explained by natural selection for both DNA and RNA viruses respectively [15]. Similarity in the codon usage may also reflect similarity in the virus life cycle and evolutionary pressure that imposes on the virus. [16] Demonstrated a positive correlation in codon usage preferences among RNA viruses that target the same host category, such as viruses infecting vertebrate hosts have different codon usage preferences to those of invertebrate viruses [16]. Recently, codon usage pattern in the genomes of SAR-CoV-2 has been analyzed. The studies suggested that the SAR-CoV-2 genomes are predominantly shaped by mutation pressure and that the codon bias is relatively low in these genomes [17]. In baculovirus genomes, the codon usage bias has shown to be weak and correlates with GC content, not with the gene length and expression level [18, 19]. The pattern of codon usage has also been suggested to be independent of the insect-host species and that natural selection dominates mutation in shaping the codon usage in baculoviruses [19]. In this study, we would like to employ Principle Component Analysis (PCA), using Relative Synonymous Codon Usage (RSCU) indices, to investigate the codon usage patterns of nucleopolyhedroviruses. We analyzed all Open-Reading-Frames (ORFs) in the genomes of 6 alphabaculoviruses; *Adoxophyes honmai* NPV, *Agrotis ipsilon* MNPV, *Autographa californica* MNPV, *Bombyx mori* NPV, *Epiphyas postvittana* NPV, *Helicoverpa armigera* NPV and 1 nudivirus (*Penaeus monodon* nudivirus). Since mutation and natural selection affect the codon usage as well as the GC content, we hypothesize that the Open Reading Frames (ORFs) that are subjected to the same type and degree of evolutionary forces should have similar pattern of RSCU, and thus cluster at the same position on the PCA plot. Nucleotide composition and G+C content at the three codon positions were analyzed to determine the type of evolutionary force that influence the codon usage of alphabaculovirus genomes and genes. In addition, we explored a gene that appear to have a unique pattern of RSCU and showed an evidence of natural selection that strongly acts on this gene.

Material and Methods

Genome Sequence and Analysis

Complete genome sequences of 6 alphabaculoviruses and nudivirus were obtained from the Gen- Bank database (http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html). *Penaeus monodon* nudivirus is used to represent a virus from a different family (*Nudiviridae*) that also produces an occlusion body similar to baculoviruses. This virus is a causative agent of spherical baculovirosis in shrimp (*Penaeus monodon*) [20]. The total of 950 ORFs were used for calculating RSCU indices (Table 1) [20-25]. The nucleotide analyses were also performed using the CAIcal server (http://genomes.urv.cat/CAIcal/) [26]. The results of the nucleotide composition analyses are in the Supplementary data 1. Sequences of insect protamine-like genes belonging to insects of the order *Lepidoptera*; *Pieris rapae* (Accession: XM_022257694.2), *Pa- pilio machaon* (Accession: XM_014512004.2), *Amyelois transitella* (Accession: XM_013327832.1) and *Helicoverpa armigera* (Accession: XM_021342924.1), were also obtained from the GenBank database. The statistical tests were carried out by the Prism 9® program. Multiple sequence align- ments were performed using COBALT (https://www.ncbi.nlm.nih.gov/guide/homology/).

Measures of Relative Synonymous Codon Usage (RSCU)

The Relative Synonymous Codon Usage (RSCU) represents the frequency for which the codon is used relative to other synonymous codons, thus providing a metric for determining whether a muta- tion replaces a more common codon with a rarer codon or vice versa [27]. The relative synonymous codon usage (RSCU) is significant to the analysis of codon bias in terms of fre- quency. An important advantage of this index is its independence from amino acid composition bias. We use the CAIcal server to calculate the RSCU (http://genomes.urv.es/CAIcal/). The RSCU value of each codon was calculated as follows:

$$RSCU = \frac{g_{ij}}{\sum_{j}^{ni} g_{ij}} ni$$

Where the value is the observed number of the gth codon for the jth amino acid which has kinds of synonymous codons. Codons with higher (or lower) selected frequencies have higher (or lower) RSCU values. Hence, a frequent codon will have an RSCU > 1 and codons with RSCU < 1 are qualified as rare, which are the characteristics of a bias codon preference. The RSCU data of 950 ORFs is in the Supplementary data 2.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was carried out using the BioVinci® program. The greatest variance represented by any projection of the data lies on the first coordinate, the so called first principal component (PC1), the second greatest variance lies on the second PC (PC2), and so on. To minimize the effect of amino acid composition on codon usage, each coding sequence was represented as a 59 dimensional vector, and each dimension corresponds to the RSCU value of each codon, which only includes synonymous codons for a particular amino acid excluding the codons AUG, UGG, and the three stop codons.

Results

Nucleotide composition analyses

Means of the overall nucleotide composition and the nucleotide composition at the three codon po- sitions of the ORFs belonging to AdhoNPV, AgipMNPV, AcMNPV, BmNPV, EppoNPV, HearNPV and PmNV genomes were analysed. In all genomes, except that of AgipMNPV, are AT-rich genome with the %A+T ranges 58.5% - 65.2% and the %G+C ranges from 34.8% - 41.5%, in which the ORFs of PmNV contains the highest %A+T and the lowest %G+C contents (Figure 1A). The genome of AgipMNPV shows similar %A+T and %G+C contents at 50.93% and 49.08%, respectively. The means of nucleotide compositions at the three codon positions reveal that the highest %A+T content is found at the second codon position for all genomes, ranging between 65%-69.7%, which is significantly different to the

Table 1: Genome sizes and number of ORFs of the 6 baculoviruses and 1 nudivirus, and their fami- ly of hosts, which all belong to the order *Lepidoptera*.

Virus	Genome Size(bp)	No. ORFs	Family of hosts	Accession no.
Adoxophyes honmai NPV (AdhoNPV)	113,220	125	Tortricidae	NC_004690
Agrotis ipsilon MNPV (AgipMNVP)	155,122	163	Noctuidae	NC_011345
Autographa californica MNPV (AcMNPV)	133,894	155	Noctuidae	NC_001623
Bombyx mori NPV (BmNPV)	128,413	143	Bombycidae	NC_001962
Epiphyas postvittana NPV (EppoNPV)	118,584	136	Tortricidae	AY043265
Helicoverpa armigera NPV (Hear-NPV)	136,740	113	Noctuidae	KJ909666
Penaeus monodon nudivirus (PmNV)	119,638	115	Panaeidae (Shrimp)	NC_024692

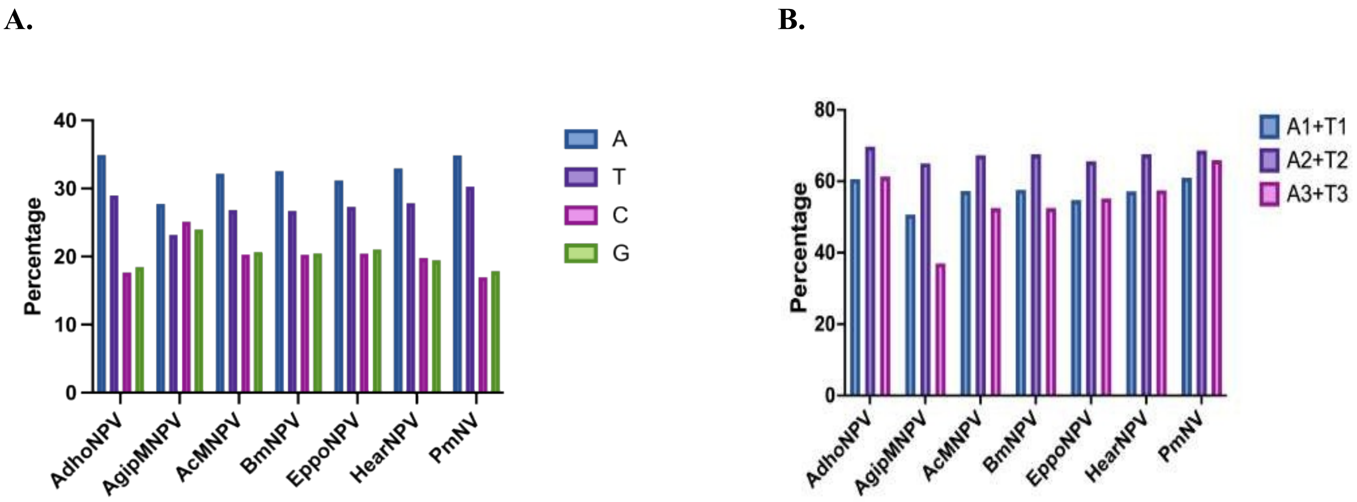


Figure 1: Analyses of nucleotide compositions. (A) means of the nucleotide composition of all ORFs from each of the alphabaculoviruses and a nudivirus, (B) means of the nucleotide composi- tions at the three codon positions.

%A+T contents at the first and third codon positions (P-value < 0.005) (Figure 1B). The %A1+T1 and %A3+T3 contents are comparable with P-value > 0.05, ranging between 50.7% - 61% and 37% - 65.9%, respectively. The lowest %A1+T1 and %A3+T3 are found in the AgipMNPV genome.

Principal Component Analysis of the RSCU indices

The RSCU indices of the 950 ORFs, belonging to the 6 alphabaculoviruses and 1 nudivirus were calculated and subjected to the Principal Component Analysis (PCA). The results shows that both Principal Components (PC1 and PC2) explains 70% of the data variation among the 59 RSCU in- dices, where PC1 and PC2 accounts for 64.87% and 5.14% of the data variance, respectively (Table 2). The ORFs clusters into 3 distinct groups; cluster 1 represents the ORFs from AgipMNPV genome, cluster 2 represents the ORFs of PmNV and cluster 3 represents the ORFs of the rest of the alphabaculoviruses (Figure 2). These results suggest that the ORFs of all alphabaculoviruses investigated, except those of AgipMNPV share similar patterns of RSCU indices. Correlation analysis (P-value < 0.05) shows that PC1 has the highest positive correlation with codon TTG (r = 0.42), followed by GAA (r = 0.35), GGC (r = 0.34), GTG (r = 0.33), AAA (r = 0.33) and the negative cor- relation with AGG (r = -0.24), followed by TCA (r = -0.23) and CTA (r = -0.22). While the PC2 has positive correlation with AGA (r = 0.59), followed by TTA (r = 0.54), GTA (r = 0.52) and negative correlation with GGC (r = -0.62), CTG (r = -0.53) and ATC (r = -0.52).

Table 2: Statistical Test of PCA

Component	Standard Deviation	Proportion of Variance
PC1	6.186842999	0.64876
PC2	1.741881609	0.05143

The cluster distribution is influenced by variations in the PC2, and all the codons with the positive correlation have A at the third codon position, while those with the negative correlation have G/C at the third codon position. In addition, the clusters appear to arrange according to the nucleotide composition, where the upper cluster of AgipMNPV ORFs has the highest %G+C and %G3+C3 (49.08% and 63%, respectively), the lowest cluster of PmNV ORFs has the lowest %G+C and %G3+C3 (34.83% and 34.1%, respec- tively), and the middle cluster has the %G+C and %G3+C3 in the middle range. There are also some ORFs that do not

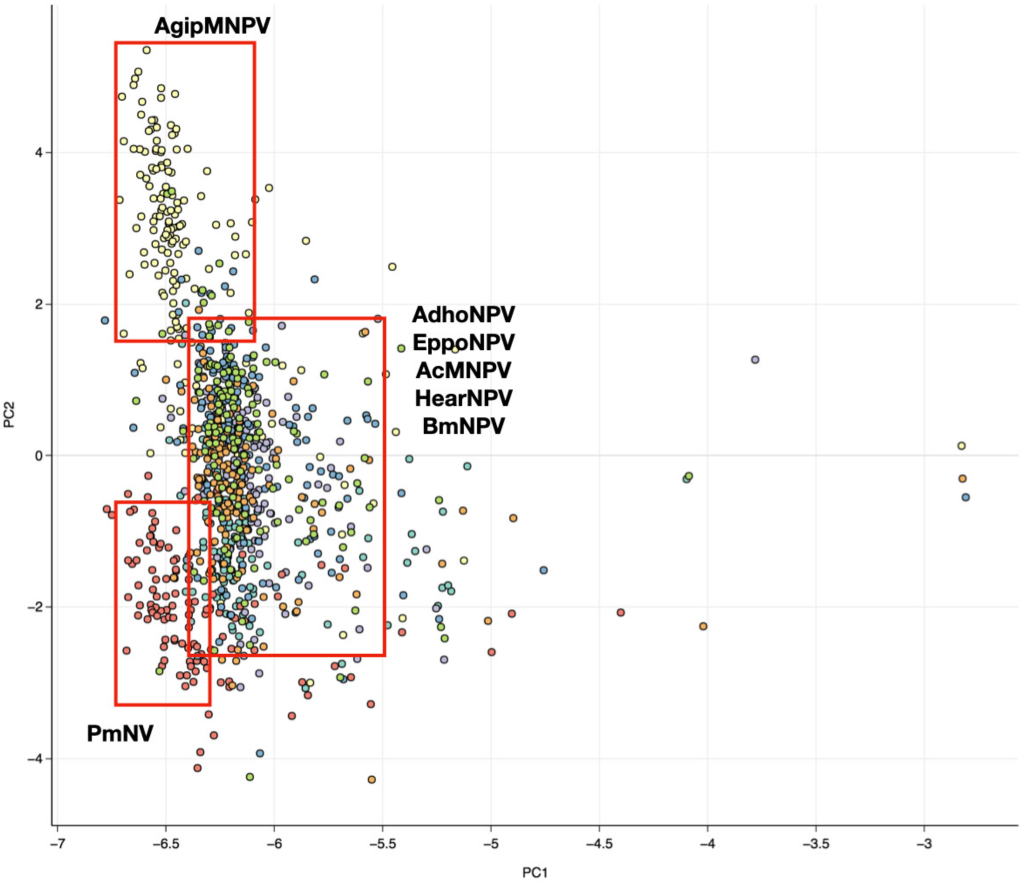


Figure 2: Two-dimensional PCA plot of the ORFs from six alphabaculoviruses; Adho- NPV (mint green), AgipMNPV (yellow), EppoNPV (grey), AcMNPV (blue), HearN- PV (orange), BmNPV (light green), and one nudivirus, PmNV (red).

cluster and position in the far-right area of the plot. Those outliers have been identified as either hypothetical or p6.9 genes. The one-dimensional PCA of all ORFs were plotted to further emphasise on the position of the outliers (Figure 3). Interestingly, one outlier of the nudivirus has also been identified as p6.9 gene. Since the p6.9 genes from all the genomes are shown to be the outliers in the PCA plot, suggesting that the gene has a different RSCU index compared to those of the rest of the ORFs. So the p6.9 genes from all alphabaculovirus genomes are subjected for further analyses. In addition, we explore the relationship between the RSCU indices and insect- host specificity, but our results show no clear relationship. The ORFs of AcMNPV and BmNPV that infect 2 different families of insect hosts show a tight clustering pattern, while the ORFs of AcMNPV and AgipMNPV that infect the same family of insect show two distinct clusterings (see Sup- plementary data 3).

The G+C content of p6.9 gene homologs

The overall %G+C content of the p6.9 gene homologs are high, ranging from 56 to 67%, which are higher than the means %G+C content of the genomes (Table 3). The gene helicase is one of the core genes that locates in the main clusters so it was selected for comparison with the p6.9 gene. The %G+C content of helicase gene homologs from the 6 alphabaculovirus genomes ranges between 34-50%, with the highest %G+C found in AgipMNPV helicase. This result is consistent with the means %G+C of all ORFs describe above. The P-value shows that the overall %G+C of the p6.9 gene homologs are significantly different to that of the helicase gene homologs (P-value = 0.0002). The %G+C at the three codon positions in the p6.9 gene homologs show an interesting pattern, by which the %G2+C2 exhibits a significantly high value, ranging between 80 and 94%. In contrast, the helicase %G2+C2 is low, ranging between 25%-30%, which is consistent with the means %G2+C2 of all ORFs.

The baculovirus p6.9 gene has been annotated as a protamine-like gene, in which its homologs in the insect host genomes are also found. We thus explored sequence relationships, focusing on the G+C content, between the baculovirus p6.9 and the host insect protamine- like genes (Table 3). Four different genera of insects belonging to the order Lepidoptera

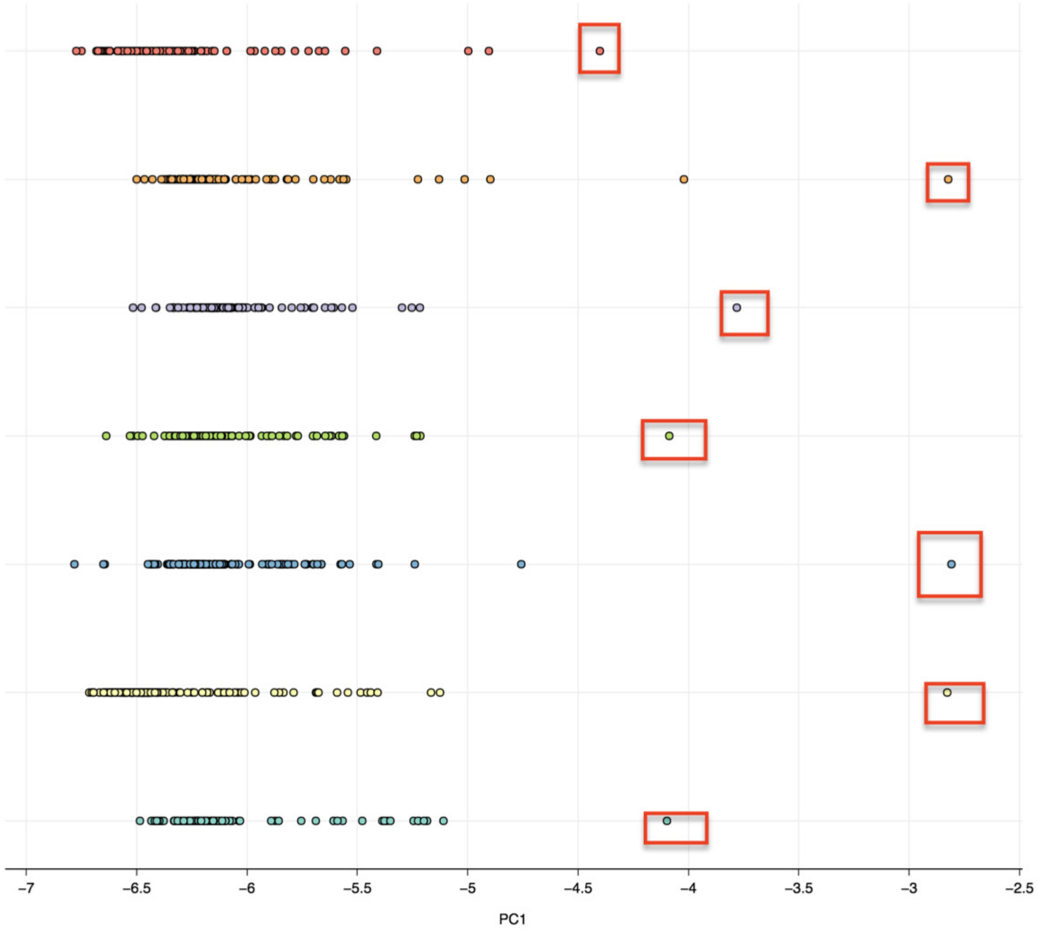


Figure 3: One-dimensional PCA plot of ORFs from all seven viruses. AdhoNPV (mint green), AgipMNPV (yellow), EppoN- PV (grey), AcMNPV (blue), HearNPV (or- ange), BmNPV (light green), and one nudi- virus, PmNV (red). The *p6.9* genes (outlier) are in red squares.

were selected for comparison. The insect protamine-like genes also exhibits a high %G+C, ranging between 55-68%, and shows no difference to that of the alphabaculovirus p6.9 genes with P-value = 0.5. The %G+C at the three codon positions are consistently higher than 50%. The alignment of the p6.9 genes and protamine-like genes are shown in the Supplementary data 4.

Sequence and Codon usage of p6.9 gene homologs

The gene *p6.9* is one of the core genes that is present in all genera of baculoviruses and it is the shortest core gene. The length of the amino acid sequences vary depending on the species of the viruses, such as EppoNPV *p6.9* gene has only 52 amino acids, while HearNPV *p6.9* gene has 110 amino acids. The *p6.9* genes use a few different species of amino acids ranging from 8-12 different species (Table 4). All sequences are arginine-rich, in which this amino acid contributes to approxi- mately 35%-44% of the sequences, followed by serine present between 12% and 23% of the sequences. Sequences of HearNPV and AgipMNPV also have a high percentage of glycine, at 32% and 20% respectively. The amino acid sequence alignment shows an evidence of either deletions or insertions, which are indicated by the alignment gaps (Figure 4). Amino acids arginine, serine, proline, threonine, phenylalanine and glycine are conserved in all p6.9 sequences. For codon usage, the p6.9 gene homologs use between 19-28 different codons, which is less than 50% of the 59 different codons available. This is partially due to a very limited usage of amino acids. Rare amino acids, such as those that are not conserved across the six p6.9 sequences, often use 1 synonymous codon, thus exhibit high RSCU value, such as TTA is used exclusively for a Leucine in the AdhoNPV (RSCU = 6), GTC for Valine in EppoNVP (RSCU = 4) and GCC for Alanine in BmNPV (RSCU = 4) (Table 5). Conserved amino acids use almost the entire set of synonymous codons. The majority of the codons exhibits RSCU ≥ 1. Overall these results suggest a degree of codon usage bias in p6.9 gene.

Discussion

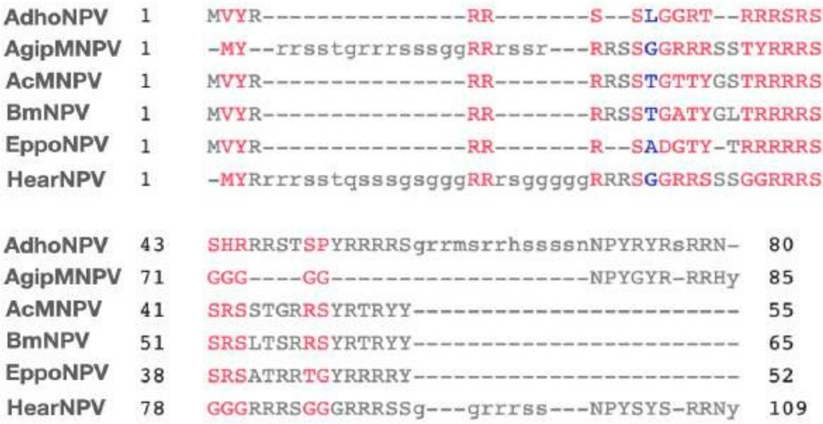
The overall genomic nucleotide composition of the 5 alphabaculoviruses (AdhoNPV, EppoNPV, AcMNPV, HearNPV and BmNPV), except AgipMNPV, suggests that the alphabaculoviruses prefer AT-rich genome. This observation is consistent with the nucleotide compositions at the three codon positions that exhibit higher A/T than G/C contents , with the highest A/T content at the second codon position. Correlation between the overall nucleotide composition and the nucleotide composition at the third codon position is a reflection of a small influence of mutational bias on the genome. Mutations at the third codon position is subjected to the codon redundancy and wobble pairing, any changes at this position do not affect the amino acid coded. The mutational bias towards A/T is perhaps due to a higher rate of G/C to A/T transitions. The higher genomic G+C content in the AgipMNPV genome is likely influenced by the G/C content at the third codon position, which may suggest that the rate of A/T to G/C transition in this genome is higher

Table 3: Overall % G+C content and the %G+C at the three codon positions

Organisms	Genes	Overall % G+C	%G1+C 1	%G2+C2	%G3+C3
AdhoNPV	P6.9	57	34	**80**	60
EppoNPV	P6.9	63	55	**82**	57
AgipMNPV	P6.9	67	51	**89**	63
AcMNPV	P6.9	56	41	**85**	46
HearNPV	P6.9	64	57	**94**	44
BmNPV	P6.9	56	44	**83**	44
AdhoNPV	Helicase	34	36	25	42
EppoNPV	Helicase	39	41	30	47
AgipMNPV	Helicase	50	45	30	75
AcMNPV	Helicase	41	39	28	55
HearNPV	Helicase	37	40	30	42
BmNPV	Helicase	39	38	27	52
Pieris rapae (white and yellow butterfly)	Protamine-like	56	55	62	51
Papilio machaon (Swallow tail butterfly)	Protamine-like	68	56	56	92
Amyelois transitella (monotypic snout moth)	Protamine-like	68	63	66	74
Helicoverpa armigera (Cotton bollworm)	Protamine 2-like	55	54	53	59

Table 4: Summary of a number of amino acid and codon usage in the alphabaculovirus *p6.9* gene homologs

Baculoviruses	No. of different amino	No. of codons with	No. of codons with
	acids and codons	RSCU ≥ 1	RSCU < 1
AdhoNPV	11, 28	18	10
EppoNPV	10, 22	19	3
AgipMNPV	8, 23	13	10
AcMNPV	8, 19	13	6
HearNPV	8, 21	15	6
BmNPV	10, 22	15	7



```
AdhoNPV   1   MVYR---------------RR-------S--SLGGRT--RRRSRST----S--RTRRRSSYYKRRPGRPR------KRSG  42
AgipMNPV  1   -MY--rrsstgrrrsssggRRrssr---RRSSGGRRRSSTYRRRSS-ggrRRSGSRRRSSGYHRRPGRPRRS---rRRSG  70
AcMNPV    1   MVYR---------------RR-------RRSSTGTTYGSTRRRRSS--------------GYRRRPGRPRTY----RRSR  40
BmNPV     1   MVYR---------------RR-------RRSSTGATYGLTRRRRSS----AGITRRRRSSGYRRRPGRPRTY----RRSR  50
EppoNPV   1   MVYR---------------RR-------R--SADGTY-TRRRRRSS--------------GYKRRPGRPRTY----RRSR  37
HearNPV   1   -MYRrrrsstqsssgsgggRRrsgggggRRRSGGRRSSSGGRRRSSsgggRRGGGRRRSGG--GGGGRRRRSsggrRRSG  77

AdhoNPV   43  SHRRRSTSPYRRRRSgrrmsrrhssssnNPYRYRsRRN-  80
AgipMNPV  71  GGG----GG-------------------NPYGYR-RRHy  85
AcMNPV    41  SRSSTGRRSYRTRYY------------------------  55
BmNPV     51  SRSLTSRRSYRTRYY------------------------  65
EppoNPV   38  SRSATRRTGYRRRRY------------------------  52
HearNPV   78  GGGRRRSGGGRRRSSg---grrrss---NPYSYS-RRNy  109
```

Figure 4: Sequence alignment of the *p6.9* gene homologs from the genomes of AdhoNPV, EppoN- PV, AgipMNPV, AcMNPV, HearNPV and BmNPV

or the rate of G/C to A/T is lower compared to the other genomes. The dominant evolutionary force that acts on the alphabaculovirus genomes is natural selection because the AT-rich genomes of the alphabaculoviruses appear to be maintained by the A/T content at the second codon position. Moreover, the AgipMNPV genome also maintains a high A/T content at this position. This suggests a higher influence of natural selection on the genomes since changes at this position affect the coding amino acid and thus protein function. Natural selection acts on the nucleotide content of a genome when the percentage of A/T or G/C affects its fitness and survival. For example, Auewarakul 2004 showed that the G/C content directly affects the viral codon adaptation index and codon usage preference, which plays a key role in predicting the efficiency of viral gene expression in the host cells [28]. The G/C content also plays an important role in the adaptation to the host environ- ment as shown in the study by [29] that Herpes Simplex Virus-1 (HSV-1) uses its high G/C content to protect itself from the insertion of an AT-rich retrotransposon (L1) abundantly found in the brain [29].

Principle Component Analysis (PCA) has shown that the ORFs from the genomes of the 5 alphabaculoviruses, AdhoNPV, EppoNPV, AcMNPV, HearNPV and BmNPV cluster at a similar position, while those of PmNV and AgipMNPV genomes form another 2 distinct clusters. This sug gests that the ORFs of the 5 alphabaculoviruses genomes possess similar codon usage pattern, while the ORFs of the other two possess different patterns, reflecting the type and degree of evolutionary pressure that act on these ORFs. Our result also shows that the clusters were arranged according to the G+C content, consistent with previous study by Jiang et al. 2008, suggesting the codon usage pattern correlates with the G+C content [9]. The cluster of the 5 al- phabaculovirus ORFs supports our hypothesis that genomes of the related viruses should possess similar codon usage pattern because they share similar mechanisms of virus replication and infection, thus are subjected to the same type of evolutionary pressure. The different pattern of the codon usage in the AgipMNPV genome is influenced by the high %G+C and %G3+C3 and perhaps implied a divergent evolution that results from a variation in the rate of transitional mutation from A/T to G/C. In addition, there is no obvious correlation between the codon usage pattern and the insect- host specificity as shown by the fact that the ORFs of AcMNPV and BmNPV that infect 2 different families of insect hosts show a tight clustering pattern, while the ORFs of AcMNPV and AgipMNPV that infect the same family of insects show two distinct clusterings. This result is consistent with previous study by [19] showing that insect-host specificity has no influence on the usage of codon in baculoviruses [19]. We further looked at the codon usage pattern of the outliers that appear in the PCA plots from all the genomes tested, by which they have been identified as

Mahapattanakul P and Wongkamhang PR., Arch Microbiol Immunology 2022
DOI:10.26502/ami.93650095

Table 5: RSCU indices of *p6.9* gene homologs from the genomes of 6 alphabaculoviruses

	Amino acids	Codons	AdhoN- PV	EppoN- PV	AgipMN- PV	AcMNPV	HearNPV	BmNPV
1	L	TTA	6	0	0	0	0	0
	L	TTG	0	0	0	0	0	6
2	I	ATC	0	0	0	0	0	3
3	V	GTT	0	0	0	4	0	4
	V	GTC	0	4	0	0	0	0
	V	GTA	4	0	0	0	0	0
4	S	TCT	0.667	1	0.316	0.6	0.96	1.091
	S	TCC	0.333	1	0.316	0.6	0.24	0.545
	S	TCA	0	1	0.316	1.2	0.72	1.636
	S	TCG	0.667	0	0.947	1.8	1.2	0.545
	S	AGT	1.333	0	0.947	0.6	0.72	0.545
	S	AGC	3	3	3.16	1.2	2.16	1.636
5	P	CCT	1	2	0	0	0	2
	P	CCC	1	0	2.67	2	4	0
	P	CCA	0	0	0	0	0	0
	P	CCG	2	2	1.33	2	0	2
6	T	ACT	1	0.8	4	0	4	0
	T	ACC	0	0.8	0	1.143	0	0.57
	T	ACA	0	0	0	2.857	0	3.429
	T	ACG	3	2.4	0	0	0	0
7	A	GCC	0	2	0	0	0	4
	A	GCA	0	2	0	0	0	0
8	Y	TAT	0.333	0	1	0.857	1	0.857
	Y	TAC	1.667	2	1	1.143	1	1.143
9	H	CAT	0	0	1	0	0	0
	H	CAC	2	0	1	0	0	0
10	Q	CAA	0	0	0	0	2	0
11	N	AAT	0.667	0	2	0	1	0
	N	AAC	1.333	0	0	0	1	0
12	K	AAA	1	2	0	0	0	0
	K	AAG	1	0	0	0	0	0
13	D	GAC	0	2	0	0	0	0
14	R	CGT	0.75	1.043	0.88	1.091	1.105	1.385
	R	CGC	1.125	1.826	1.94	2.455	1.105	2.308
	R	CGA	0.375	1.826	0.71	0.273	1.421	0.231
	R	CGG	0.562	0	0.18	0	0	0
	R	AGA	2.062	1.043	1.24	1.636	1.737	1.615
	R	AGG	1.125	0.261	1.06	0.545	0.632	0.231
15	G	GGT	0.8	0	0.47	4	1.556	4
	G	GGC	2.4	3	3.39	0	2	0
	G	GGA	0.8	1	0.24	0	0.444	0
	G	GGG	0	0	0	0	0	0

a protamine-like genes (*p6.9*). It is interesting that the *p6.9* gene is also an outlier in the nudivirus ORF cluster, which is from a different family of virus. The distinct RSCU index of the *p6.9* gene is perhaps a reflection of its exposure to a different evolutionary pressure and distinct function in the occlusion-forming viruses.

The overall G+C content of the alphabaculovirus *p6.9* genes shows a distinct pattern compared to the *helicase* genes that represent genes from the main cluster. The *p6.9* genes have a significantly higher G+C content when compared to the *helicase* genes, while there is no significant difference in the G+C content between the *p6.9* and the insect protamine-like genes. This may reflect a functional relationship between these protamine-like genes. The high %G2+C2 in the *p6.9* gene is extremely significant, and this coincides with the high percentage of arginine, serine and glycine in the sequences. These amino acids have G or C at their second codon position. Thus, the abundance of these 3 amino acids in the *p6.9* sequences contributes to the high percentage of the gene G+C content, especially at the second codon position. A high content of arginine and its positively-charged property, which is also known to be a characteristic of a protamine gene, has been selected for its ability to compact DNA to a very high density [30, 31]. Tight DNA packing has also been proposed to prevent DNA damage from radicals, as well as to inactivate the gene. Therefore, the high %G2+C2 in the *p6.9* gene is likely to reflect its function in DNA packaging. This is an evidence of a strong natural selection that acts on a gene, effecting the codon usage and the nucleotide composition of the gene. Comparing to the insect protamine like genes, the %G2+C2 content is much higher in the *p6.9* genes. This may owe to the fact that the insect pro- tamine-like genes have adapted to the insect host and use a more diverse set of amino acids that contributes to the %A+T at the second codon position. The *p6.9* genes use a few species of amino acids, perhaps reflecting an adaptation to the virus lifestyle that is dependent on host insects. The *p6.9* gene homologs exhibit a similar pattern in the codon usage preference, by which almost the entire sets of synonymous codons are used to encode conserved amino acids. While only one syn onymous codon is used to encode rare amino acids. Since the conserved amino acids are present in all *p6.9* sequences, this indicates that these amino acids play an important role in the function of the P6.9 protein. Usage of the entire set of codons is perhaps a mechanism to ensure that these amino acids will get translated effectively. In contrast, the rare amino acids are not conserved among different species, suggesting that they are not important. This is perhaps an evidence of genetic drift, by which the codons have changed or been replaced because they play no role in maintaining the function of the protein.

In conclusion, we have shown that the predominant evolutionary force that influences the alphabaculovirus AT-rich genomes is natural selection, by which the natural selection acts on the codon usage that in turn influences the overall nucleotide composition. Principle Component Analysis based on RSCU indices can serve as a tool to distinguish genes that are subjected to similar or different types and degree of evolutionary force as shown by the clustering pattern and the presence of the outliers in the PCA plot. Overall, ORFs that belong to closely related organisms tend to expose to similar evolutionary pressure and thus exhibit similar codon usage pattern. In addition, the distinct RSCU index of *p6.9* genes demonstrates that a gene within the same genome may be under to a different type and degree of selection pressure that results in a distinct pattern of codon usage.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests

Funding

No funding was received for conducting this study.

Authors' contributions:

P. Mahapattanakul acquired the sequence data, performed the RSCU, PCA and nucleotide composition analyses, Interpret the results and draft the manuscript. P. Rodpothong designed the work, performed the *p6.9* sequence analyses, Interpret the results and finalize the manuscript.

Supplementary Information

Click here:

References

1. Rohrmann GF. Baculovirus Molecular Biology. th Bethesda (MD) (2019).
2. Clem RJ and Passarelli AL. Baculoviruses: sophisticated pathogens of insects. PLoS Pathog 9 (2013): e1003729.
3. Sajjan DB and Hinchigeri SB. Structural Organization of Baculovirus Occlusion Bodies and Protective Role of Multilayered Polyhedron Envelope Protein. Food Environ Virol 8 (2016): 86-100.
4. Herniou EA, Arif BM and Becnel JJ. Family Baculoviridae. Virus Taxonomy, Ninth Report of the International Committee on Taxonomy of Viruses. AMQ King, MJ Adams, EB Carstens and EJ Lefkowitz. Amsterdam, Elsevier Academic Press (2012): 163-173.

5. Herniou EA, Olszewski JA, Cory JS and O'Reilly DR. The genome sequence and evo- lution of baculoviruses. Annu Rev Entomol 48 (2003): 211-234.

6. Herniou EA and Jehle JA. Baculovirus phylogeny and evolution. Curr Drug Targets 8 (2007): 1043-1050.

7. van Oers MM and Vlak JM. Baculovirus genomics. Curr Drug Targets 8 (2007): 1051-1068.

8. Miele SA, Garavaglia MJ, Belaich MN and Ghiringhelli PD. Baculovirus: molecular insights on their diversity and conservation. Int J Evol Biol (2011): 379424.

9. Jiang Y, Deng F, Wang H and Hu Z. An extensive analysis on the global codon usage pat- tern of baculoviruses. Arch Virol 153 (2008): 2273-2282.

10. Szewczyk B, Rabalski L, Krol E, Sihler W and de Souza ML. Baculovirus biopesticides – a safe alternative to chemical protection of plants. Journal of Biopesticides 2 (2009): 209-216.

11. Kost TA, Condreay JP and Jarvis DL. Baculovirus as versatile vectors for protein ex- pression in insect and mammalian cells. Nature Biotechnology 23 (2005): 567-575.

12. Hitchman RB, Possee RD and King LA. Baculovirus expression systems for recombi- nant protein production in insect cells. Recent Pat Biotechnol 3 (2009): 46-54.

13. Shackelton LA, Parrish CR and Holmes EC. Evolutionary basis of codon usage and nu- cleotide composition bias in vertebrate DNA viruses. Journal of Molecular Evolution 62 (2006): 551-563.

14. Jenkins GM and Holmes EC. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus research 92 (2003): 1-7.

15. Chen Y. A comparison of synonymous codon usage bias patterns in DNA and RNA virus genomes: quantifying the relative importance of mutational pressure and natural selection. Biomed Res Int (2013): 406342.

16. Su MW, Lin HM, Yuan HS and Chu WC. Categorizing host-dependent RNA viruses by principal component analysis of their codon usage preferences. J Comput Biol 16 (2009): 1539-1547.

17. Khattak S, Rauf MA, Zaman Q, Ali Y, Fatima S, Muhammad P, et al. Genome-Wide Analysis of Codon Usage Patterns of SARS-CoV-2 Virus Reveals Global Heterogeneity of COVID-19. Biomolecules 11 (2021).

18. Levin DB and Whittome B. Codon usage in nucleopolyhedroviruses. J Gen Virol 81 (2000): 2313- 2325.

19. Shi SL, Jiang YR, Yang RS, Wang Y and Qin L. Codon usage in Alphabaculovirus and Betabaculovirus hosted by the same insect species is weak, selection dominated and exhibits no more similar patterns than expected. Infect Genet Evol 44 (2016): 412-417.

20. Yang YT, Lee DY, Wang Y, Hu JM, Li WH, Leu JH, et al. The genome and occlusion bodies of marine Penaeus mon- odon nudivirus (PmNV, also known as MBV and PemoNPV) suggest that it should be assigned to a new nudivirus genus that is distinct from the terrestrial nudiviruses. BMC Genomics 15 (2014): 628.

21. Ayres MD, Howard SC, Kuzio J, Lopez-Ferber M and Possee RD. The complete DNA sequence of Autographa californica nuclear polyhedrosis virus. Virology 202 (1994): 586-605.

22. Gomi S, Majima K and Maeda S. Sequence analysis of the genome of Bombyx mori nucle- opolyhedrovirus. J Gen Virol 80 (1999): 1323-1337.

23. Hyink O, Dellow RA, Olsen MJ, Caradoc-Davies KMB, Drake K, Herniou EA, et al. Whole genome analysis of the Epiphyas postvittana nu- cleopolyhedrovirus. J Gen Virol 83 (2002): 957-971.

24. Nakai M, Goto C, Kang W, Shikata M, Luque T and Kunimi Y. Genome sequence and organization of a nucleopolyhedrovirus isolated from the smaller tea tortrix, Adoxophyes honmai. Virology 316 (2003): 171-183.

25. Noune C and Hauxwell C. Complete Genome Sequences of Seven Helicoverpa armigera SNPV-AC53-Derived Strains. Genome Announc 4 (2016).

26. Puigbo P, Bravo IG and Garcia-Vallve S. CAIcal: a combined set of tools to assess codon usage adaptation. Biol Direct 3 (2008): 38.

27. Sharp PM and Li WH. An evolutionary perspective on synonymous codon usage in uni- cellular organisms. J Mol Evol 24 (1986): 28-38.

28. Auewarakul P. Composition bias and genome polarity of RNA viruses. Virus research 109 (2005): 33-37.

29. Brown JC. High G+C Content of Herpes Simplex Virus DNA: Proposed Role in Protection Against Retrotransposon Insertion. Open Biochemistry Journal 1 (2007): 33-42.

30. Brewer LR, Corzett M and Balhorn R. Protamine-induced condensation and decondensa- tion of the same DNA molecule. Science 286 (1999): 120-123.

31. DeRouchey JB, Hoover and Rau DC. A comparison of DNA compaction by arginine and lysine peptides: a physical basis for arginine rich protamines. Biochemistry 52 (2013): 3000-3009.

32. Wang J, Hou D, Wang Q, Kuang W, Zhang L, Li J, Shen S, et al. Genome analysis of a novel Group I alphabaculovirus obtained from Oxyplax ochracea. Plos One 13 (2018): e0192279.