

RESEARCH ARTICLE

Open Access



Analysis of codon usage bias of envelope glycoprotein genes in nuclear polyhedrosis virus (NPV) and its relation to evolution

Yongchao Zhao^{1†}, Hao Zheng^{1†}, Anying Xu^{2†}, Donghua Yan¹, Zijian Jiang¹, Qi Qi¹ and Jingchen Sun^{1*}

Abstract

Background: Analysis of codon usage bias is an extremely versatile method using in furthering understanding of the genetic and evolutionary paths of species. Codon usage bias of envelope glycoprotein genes in nuclear polyhedrosis virus (NPV) has remained largely unexplored at present. Hence, the codon usage bias of NPV envelope glycoprotein was analyzed here to reveal the genetic and evolutionary relationships between different viral species in baculovirus genus.

Results: A total of 9236 codons from 18 different species of NPV of the baculovirus genera were used to perform this analysis. Glycoprotein of NPV exhibits weaker codon usage bias. Neutrality plot analysis and correlation analysis of effective number of codons (ENC) values indicate that natural selection is the main factor influencing codon usage bias, and that the impact of mutation pressure is relatively smaller. Another cluster analysis shows that the kinship or evolutionary relationships of these viral species can be divided into two broad categories despite all of these 18 species are from the same baculovirus genus.

Conclusions: There are many elements that can affect codon bias, such as the composition of amino acids, mutation pressure, natural selection, gene expression level, and etc. In the meantime, cluster analysis also illustrates that codon usage bias of virus envelope glycoprotein can serve as an effective means of evolutionary classification in baculovirus genus.

Keywords: Nuclear polyhedrosis virus, Codon usage bias, Envelope glycoprotein, Evolution

Abbreviations: A3s, A content on the third synonymous codon position; C3s, C content on the third synonymous codon position; CAI, Codon adaptation index; COA, Correspondence analysis; ENC, Effective number of codons; FA, Factor analysis; G3s, G content on the third synonymous codon position; GC1, GC content in the first position of the codons; GC12, Stands for the average value of GC content in the first and second position of the codons; GC2, GC content in the second position of the codons; GC3, GC content in the third position; GC3s, GC content on the third synonymous codon position; LRA, Linear regression analysis; RSCU, Relative synonymous codon usage; T3s, T content on the third synonymous codon position

* Correspondence: cyfz@scau.edu.cn

†Equal contributors

¹Subtropical Sericulture and Mulberry Resources Protection and Safety Engineering Research Center, Guangdong Provincial Key Laboratory of Agro-animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou 510642, People's Republic of China

Full list of author information is available at the end of the article



Background

Codons are not used equally in most organisms. During an organisms evolutionary history, preference for using a particular synonymous codon will be formed within a species or gene in the long-term. Codons which are used in higher frequency within species or genes are referred to as optimal codons. Codon usage bias itself refers to such cases in which codons are utilized in higher frequency than other synonymous codons during the process of translation, often as a result of adaptive evolution [1]. Analysis of codon usage bias is thus of vital significance in the quest to improve exogenous gene expression levels within host cells. Codon bias analysis is a common phenomenon in many species, such as *Escherichia coli* [2], *Arabidopsis thaliana* [3], *Xanthophyllomyces dendrorhous* [4], *Taenia saginata* [5], *Megalobrama amblycephala* [6], metazoans [7], and even human beings [8]. Recent studies have shown that the employment of some special synonymous codons can affect protein folding as well as errors in folding [9, 10]. Furthermore, studies have shown that the inherent links between codon usage and thus amino acids influence the protein components of cells [11]. At the same time, thoroughly understanding codon usage bias plays a central role in making accurate prediction of related gene functions.

Different genes exhibit different codon usage bias in the same genome. Mutation, natural selection, and random drift were the three major factors for species' codon usage bias [12–15]. Bioinformatics methods analyses have shown that translation selection is probably the original reason for the formation of codon usage bias. Other possible factors affecting codon usage bias among species include: gene expression level [16], gene length [8], GC content [17], recombination rate, RNA stability [18], environmental stress [19], population size [20], evolutionary age of genes [21], and so on. Codon usage bias has profound influence on genomic evolution [22]. Even within the same genome, codon usage patterns are not necessarily the same within the same gene [23].

Envelope glycoprotein is a main fatty acid acylating glycoprotein of *Bombyx mori* nuclear polyhedrosis virus (BmNPV) [24]. With pH-dependent membrane fusion activity, it can make the virus and host cell fusion. Glycoprotein gene mainly is connected by disulfide bond into the form of trimer and exists in the end of stick baculovirus, forming a typical membrane-grain structure. Research shows that the monoclonal antibody of glycoprotein can make a significant reduction in the infectious virus particles [25]. The glycoprotein gene plays a key role in the progress that baculovirus infects cells and progeny nuclear capsid effective budding release [26, 27]. Besides, silencing glycoprotein gene in

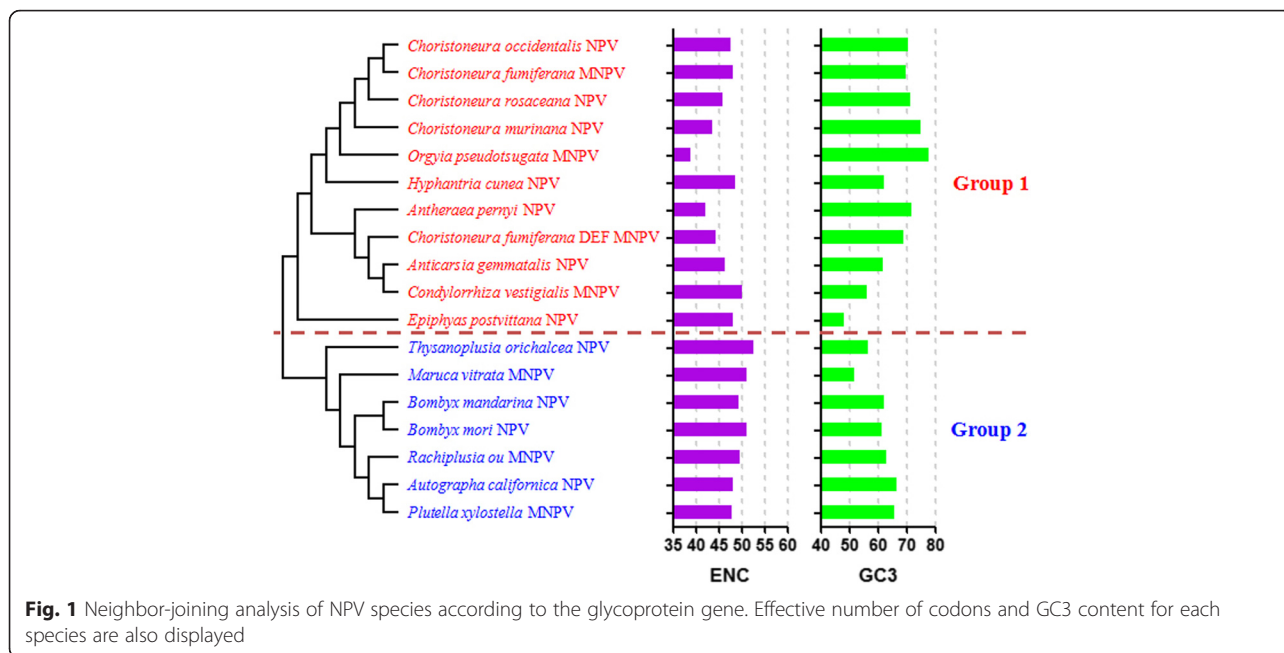
transgenic silkworm increases resistance to BmNPV [28, 29]. Therefore, the glycoprotein is one of the most important capsule membrane proteins in baculovirus. Because the glycoprotein gene product and its homologues are relatively conserved, it is an ideal gene useful in the study of the evolutionary relationships of different baculoviruses. Analyses of codon usage bias could therefore enable a better understanding of the NPV molecular evolution dynamic.

Results

Clustering analysis

A total of 18 different NPV species were analyzed using a Neighbor-joining method specifically using their glycoprotein genes for calculation. The ENC and GC content in the third position (GC3) of triplet codons from each species are compared simultaneously. It is observed that *Choristoneura occidentalis* NPV (ChocNPV), *Choristoneura fumiferana* MNPV (CfMNPV), *Choristoneura rosaceana* NPV (CrNPV), *Choristoneura murinana* NPV (CmNPV), *Orgyia pseudotsugata* MNPV (OpMNPV), *Hyphantria cunea* NPV (HycuNPV), *Antheraea pernyi* NPV (AnpeNPV), *Choristoneura fumiferana* DEF MNPV (CfDEFMNPV), *Anticarsia gemmatalis* NPV (AgNPV), *Condylorrhiza vestigialis* MNPV (CvMNPV) and *Epiphyas postvittana* NPV (EppoNPV) belong to Group 1, indicating that their evolutionary relationship is similar. But, the variation range of their ENC value and GC3 content are more extensive ($38.9 \leq \text{ENC} \leq 50.0$ and $48.4 \leq \text{GC3} \leq 77.5$, respectively) (Fig. 1). This result shows that they do not exhibit similar codon usage bias although they possess equally evolutionary position. There are 7 NPV species with *Thysanoplusia orichalcea* NPV (ToNPV), *Maruca vitrata* MNPV (MvMNPV), *Bombyx mandarina* NPV (BmaNPV), *Bombyx mori* NPV (BmNPV), *Rachiplusia ou* MNPV (RaouMNPV), *Autographa californica* NPV (AcNPV) and *Plutella xylostella* MNPV (PlxyNPV) in Group 2. The variation range of their ENC value and GC3 content are relatively smaller ($47.9 \leq \text{ENC} \leq 52.5$ and $51.9 \leq \text{GC3} \leq 66.7$, respectively). It suggests that the closer the evolution of species classification, the more similar their codon usage bias.

On the other hand, 18 different NPV species were also analyzed using a cluster analysis method specifically using RSCU values for calculation. It is observed that AcNPV, PlxyNPV, RaouMNPV, BmNPV, BmaNPV, ToNPV and MvMNPV exhibit similar evolution status, consistent with the analysis results of Neighbor-joining (Fig. 2). Other species are similar to the results of Neighbor-joining on the whole, but still have some differences from the perspective of the individual evolutionary branch, such as AgNPV, CvMNPV and HycuNPV exhibit similar codon usage bias, indicating that the respective pairs are evolutionarily related.



It shows that there are some difference between Cluster and Neighbor-joining analysis, especially codon usage patterns are rather different in Group 2 but less apparent in Group 1.

Glycoprotein codon usage bias of 18 NPVs analysis

We analyzed the glycoprotein genes of 18 NPV species. The GC content for these genes ranges from 42.7 to 54.4 %, with the average being 48.90 %. GC content varies most significantly in the first and third codon positions, with values of 46.39 and 64.44 %, respectively. The ENC in glycoprotein varies from 38.9 to 52.5, with a

mean of 47.39. Accordingly, none of the 18 glycoprotein genes exhibits strong codon bias, as all of their ENC values are above 35. This data shows that glycoprotein displays a general random codon usage, lacking strong codon bias (Table 1 and Additional file 1).

Additionally, the relative synonymous codon usage (RSCU) values of 59 sense codons (except for Trp, initiator codon and terminator codon) also support this conclusion NPV glycoprotein presenting weak codon bias. Nearly half of the glycoprotein codons (24/59) are frequently used as shown in Table 2, such as GGC (coding Glycine) and UUG (coding Leucine). The RSCU values

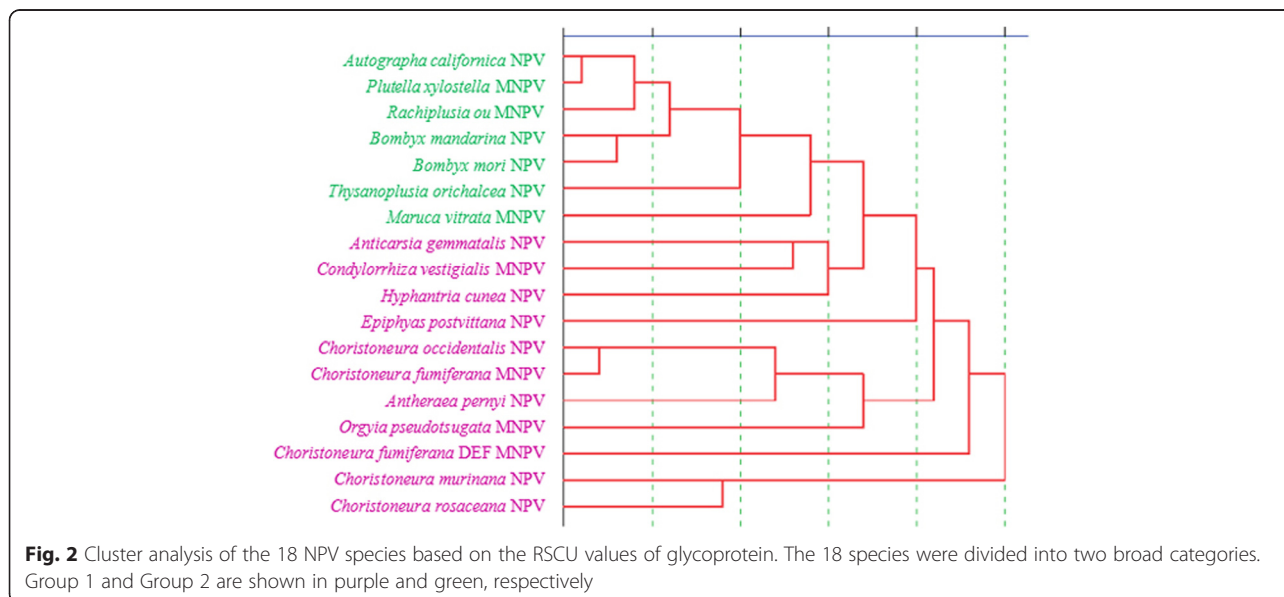


Table 1 Means and standard deviations of several index numbers from 18 glycoprotein genes in NPV

A3s (%)	T3s (%)	C3s (%)	G3s (%)	GC (%)	GC1 (%)	GC2 (%)	GC3 (%)	GC3s (%)	ENC	CAI
23.24 ± 5.01	25.24 ± 5.87	46.05 ± 5.45	37.70 ± 5.39	48.90 ± 3.13	46.39 ± 1.62	35.59 ± 1.05	64.44 ± 7.64	62.49 ± 7.98	47.39 ± 3.33	0.78 ± 0.01

of our set of NPV glycoprotein genes were analyzed, as shown in Table 2. The result was representative of NPV as a whole. All of the UUU, UUG, CUG, AUU, GUG, AGC, UCG, CCC, CCG, ACC, ACG, GCC, GCG, UAC, CAC, CAA, AAC, AAA, GAC, GAG, UGC, CGU, CGC, and GGC codons have a somewhat high bias (RSCU > 1.0), esp UUU, UUG, AUU, GUG, AGC, CGC codons (RSCU > 1.5) possess a strong bias. Other codons are used less

frequently, such as UUC, AUA, UCA, CCA, GCA, GAU, AGG, GGA, GGG, and etc. (RSCU < 1.0).

Furthermore, we also compared the RSCU values of 59 sense codons (Fig. 3). There was some difference of the RSCU values of 59 sense codons from 18 NPVs glycoprotein, but the overall trend is relatively similar. This illustrates that relatively similar species maintain the stability codon usage patterns.

Table 2 The RSCU value and used codon numbers in the glycoprotein of NPV(9236 codons)

Amino acid	Codon	Number	RSCU	Amino acid	Codon	Number	RSCU
Phe	UUU	240	1.57	Ala	GCU	72	0.68
	UUC	65	0.43		GCC	168	1.59
Leu	UUA	74	0.59	Tyr	GCA	28	0.27
	UUG	225	1.80		GCG	154	1.46
	CUU	84	0.67		UAU	84	0.59
	CUC	112	0.90	UAC	202	1.41	
	CUA	71	0.57	His	CAU	93	0.63
Ile	CUG	183	1.47	Gln	CAC	203	1.37
	AUU	336	1.65		CAA	141	1.11
	AUC	181	0.89	CAG	114	0.89	
Val	AUA	95	0.47	Asn	AAU	182	0.51
	GUU	89	0.64		AAC	528	1.49
	GUC	95	0.68	Lys	AAA	367	1.19
	GUA	95	0.68		AAG	248	0.81
Ser	GUG	278	2.00	Asp	GAU	149	0.46
	AGU	73	0.69		GAC	497	1.54
	AGC	206	1.94	Glu	GAA	286	0.97
	UCU	64	0.60		GAG	306	1.03
	UCC	81	0.76		Cys	UGU	77
UCA	42	0.39	UGC	200		1.44	
Pro	UCG	172	1.62	Arg	CGU	73	1.01
	CCU	32	0.54		CGC	196	2.72
	CCC	74	1.26		CGA	39	0.54
	CCA	22	0.37		CGG	52	0.72
	CCG	107	1.82		AGA	41	0.57
Thr	ACU	103	0.68	Gly	AGG	32	0.44
	ACC	208	1.37		GGU	79	0.67
	ACA	94	0.62		GGC	321	2.72
	ACG	204	1.34		GGA	44	0.37
				GGG	28	0.24	

Preferentially used codons are displayed in bold

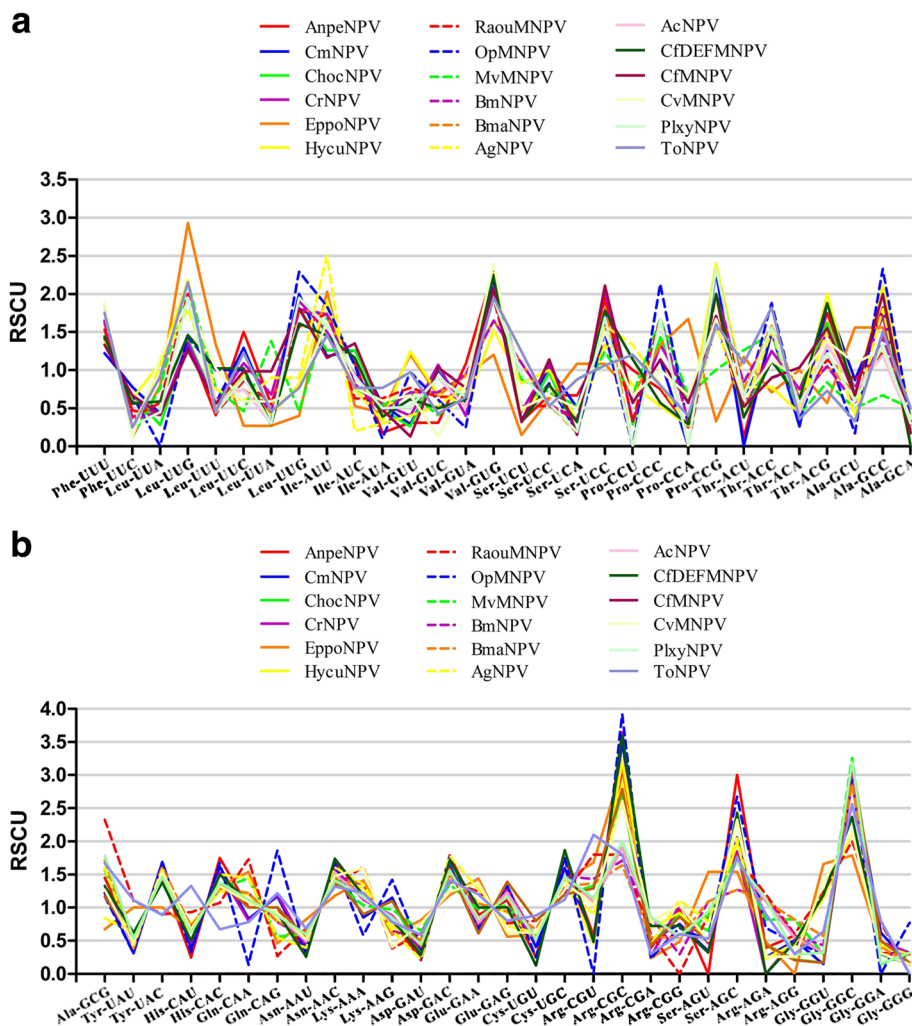


Fig. 3 Analysis of relative synonymous codon usage of glycoprotein gene in 18 NPV species. **a** and **b** show the comparison of 59 sense codons from 18 NPV glycoproteins

Nucleotide composition affects the formation of codon usage bias

Correspondence analysis was performed on the RSCU values. The axes factors, as shown in Fig. 4, are represented by Axis 1 and 2 which correlate to two main influencing factors of codon usage bias. They represent 36.29 and 21.20 % of the total variation, respectively (Fig. 4). The relationship between codon usage bias and amino acid composition were explained by multifactor variable analysis. Axis 1 has a distinct positive correlation with C3s ($r = 0.965, p < 0.01$), G3s ($r = 0.948, p < 0.01$), and GC3s ($r = 0.996, p < 0.01$). Axis 1 shows evidently negative correlation with A3s ($r = -0.957, p < 0.01$) and T3s ($r = -0.969, p < 0.01$). There is an obvious negative correlation between GC3s and ENC ($r = -0.822, p < 0.01$). However, GC3s exhibits a significantly positive correlation with Axis 1 ($r = 0.996, p < 0.01$). On the other hand, ENC shows significant negative correlation with Axis 1

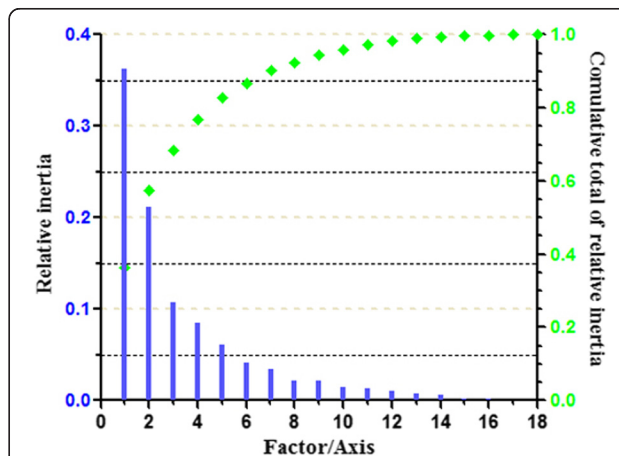


Fig. 4 The correspondence analysis of the amino acid usage frequencies in glycoprotein. The relative and cumulative inertia of the first 18 factors were shown based on the correspondence analysis

($r = -0.806, p < 0.01$) (Table 3). There is a high correlations among these parameters because their R value is greater than 0.8. These results demonstrate that nucleotide composition indeed affects codon usage bias.

All of the genes are diffusely distributed, and it indicates that many factors affect codon usage bias (Fig. 5a). Axis 1 represents the main index for affecting codon usage bias. The distribution density of triplet codons ending with G/C is closer to Axis 1 than that of codons ending with A/U (Fig. 5b). Thus, these results suggest that nucleotide composition (especially G and C) posits a certain degree of influence on the codon usage bias. Furthermore, the mutation impact of codons ending with G/C on codon usage bias is greater than that of codons ending with A/U.

GC3s affecting codon bias

A standard curve evaluates the relationship between ENC and GC3s, which illustrates their corresponding relationship of the extent under mutation pressures. If the points, which represent various genes, fall on or near the standard curve, the codon usage bias would be interpreted as being mainly determined by mutation pressures. Generally, codon usage bias depends on the content of the ending base in codons—in other words, the GC3s content of genes. However, all of the points (no matter the Group 1 or Group 2) are located beneath the standard curve, indicating that mutation pressure which is not the critical factor in the formation of codon preferences (Fig. 6). Thus, the GC3s values of glycoprotein are not the sole factors affecting codon bias formation in various species of NPV. Furthermore, the dispersed plotted genes indicates that other factors can impact codon

usage bias to a certain extent. These factors include natural selection, gene length, and gene expression levels.

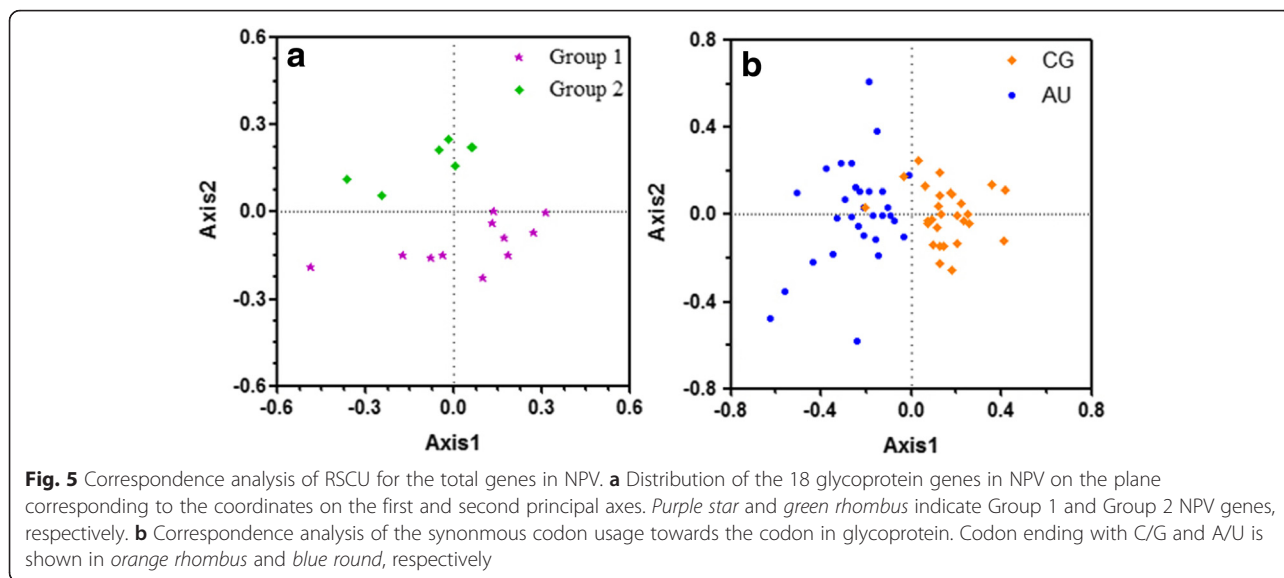
Natural selection plays an important role in the process of codon bias formation

ENC-plot analysis demonstrated the extent to which mutational pressures affect the formation of codon usage bias. We next seek to determine whether natural selection or mutation pressure plays a greater role in generating codon usage bias. To determine this, we attempt to carry out a neutrality plot analysis on the GC content of codons. The distribution range of GC3 is very broad, from 48.4 to 77.5 % (Fig. 7). There is indeed obvious correlation between GC1 and GC3 ($p < 0.01$), which initially seemed indicative of mutation pressure playing a greater role in direct codon usage bias. However, after calculating the neutrality plot, this was not the case. In Fig. 7, all of GC3 values diffuse distribution and all of regression curve deviate from the diagonal line. And then, the slope of the regression line was determined to be 0.1063, 0.0685 and 0.0956. Should the slope be equal to one (diagonal line), indicating a perfect correlation between GC12 and GC3, mutation pressure would be deemed the dominant factor in generating bias. Slopes approaching the vertical or horizontal axes would indicate natural selection as dominant. Despite the observed GC12 and GC3 correlation, our slope of 0.1063, 0.0685 and 0.0956 indicates that the influence of direct mutation pressure for codon usage bias is only 10.63, 6.85 and 9.56 %, respectively. The influence of natural selection on codon usage bias was calculated to be 89.37, 93.15 and 90.44 %, thereby indicating natural selection as the dominant factor influencing bias.

Table 3 Correlation coefficients between the position of genes along the first two major axes with index of glycoprotein genes' codon usage and synonymous codon usage bias

	Length	GC	GC1	GC2	GC3	GC3s	A3s	T3s	C3s	G3s	ENC	CAI	Axis1
GC	-0.353												
GC1	-0.498*	0.818**											
GC2	-0.672**	0.626**	0.730**										
GC3	-0.208	0.944**	0.685**	0.376									
GC3s	-0.208	0.944**	0.685**	0.376	1.000**								
A3s	0.207	-0.901**	-0.601**	-0.334	-0.959**	-0.959**							
T3s	0.283	-0.961**	-0.754**	-0.455	-0.973**	-0.973**	0.913**						
C3s	-0.308	0.973**	0.771**	0.502*	0.971**	0.971**	-0.913**	-0.979**					
G3s	-0.099	0.866**	0.571**	0.211	0.946**	0.946**	-0.973**	-0.911**	0.897**				
ENC	0.491*	-0.880**	-0.722**	-0.609**	-0.822**	-0.822**	0.763**	0.858**	-0.899**	-0.709**			
CAI	0.485*	-0.229	-0.142	-0.364	-0.154	-0.154	0.183	0.209	-0.186	-0.158	0.142		
Axis1	-0.198	0.938**	0.688**	0.378	0.996**	0.996**	-0.957**	-0.969**	0.965**	0.948**	-0.806**	0.361	
Axis2	0.760**	-0.152	-0.334	-0.671**	0.007	0.007	-0.032	0.044	-0.110	0.203	0.370	0.152	0.018

** $p < 0.01$. * $p < 0.05$

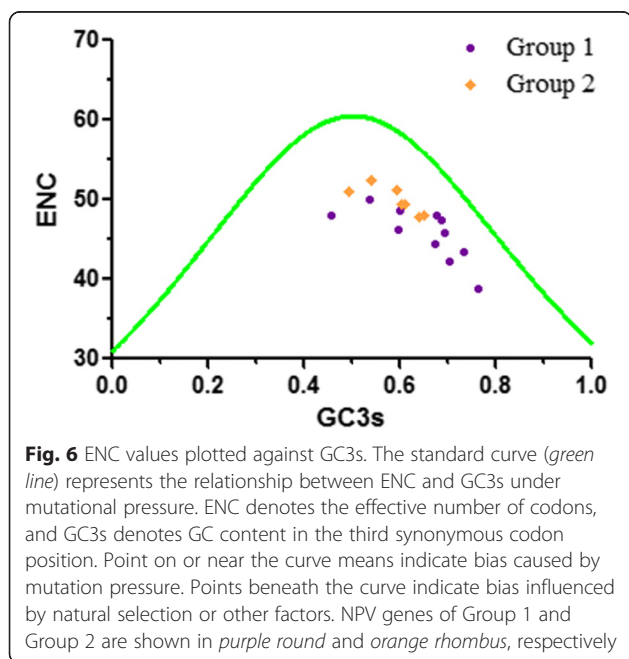


Effects of gene length and expression level on codon usage bias

CAI values are useful in predicting the levels of gene expression. Silkworm ribosomal genes, which have a high level of expression, were used as references in our computation of codon adaptation indices [30]. Correlation analysis shows that CAI and ENC demonstrate no significant correlation, as well as no obvious correlation exists between the CAI, GC3s and GC content. This illustrates that gene expression levels have no effect on codon bias. On the other hand, gene length has no obvious correlation with CAI, ENC and Axis 1. This observation indicates that there is no correlation between the

length of the gene and its codon usage bias for NPV glycoprotein.

The CAI values of the various glycoprotein genes ranges from 0.765 to 0.812, and the length of the gene ranges from 1500 to 1593 bp. The level of variation in CAI values and gene length among the various glycoprotein genes is relatively small, as shown in Fig. 8. These results indicate that gene expression level and length play an acute role in the shaping of codon bias. The gene lengths of the various viral species are all relatively similar, and given that all species CAI values are very approximate. It was suggested that the glycoprotein gene displays stable expression in the process of evolution in NPV.



Discussion

Our clustering analysis statistics are similar to Neighbor-joining. We compared the RSCU values of glycoprotein from 18 species, the results show that they have relative similarity codon usage bias. After a series of analyses, glycoprotein possess a general codon usage pattern because all the ENC values are greater than 35. RSCU values are an index for assessing frequency of synonymous codon usage. RSCU = 1.0 means that there is only one codon within a synonymous codon set, and it indicates that the codon is not biased. Alternatively, RSCU > 1.0 indicates a high frequency bias for a particular codon within a synonymous codon set, and vice versa [31]. Many factors can result in the synonymous codon usage bias of the glycoprotein gene of the NPV genus. Nucleotide composition is one of the factors that affect codon usage bias especially codons ending with G/C. But, most of the codons ending with A/U also demonstrated a stronger frequency of codon usage bias in other species that contains rich A/T base pairs, such as *Saccharomyces cerevisiae* and *Plasmodium falciparum* [32, 33]. In

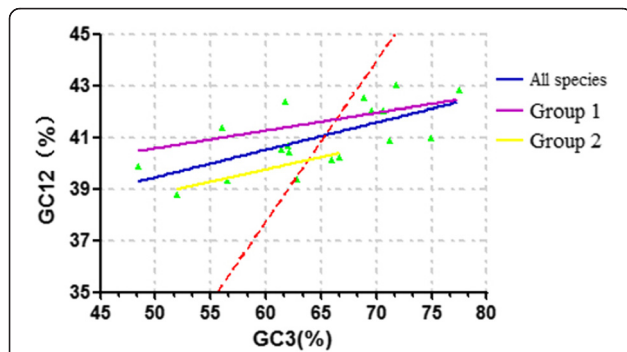


Fig. 7 Neutrality plot analysis of the 18 glycoprotein genes. Neutrality plot analysis of the average GC content in the first and second positions of the codons (GC1 and GC2) and of the GC content in the third position (GC3) for glycoprotein. The blue, purple and yellow regression curve represented as $y = 0.1063x + 34.14$, $R^2 = 0.4307$, $y = 0.0685x + 37.15$, $R^2 = 0.3517$, $y = 0.0956x + 34.02$, $R^2 = 0.5645$. The diagonal line is colored in red dotted line

addition, previous studies have identified mutational pressures and natural selection as two major factors influencing codon usage bias [34]. The ENC-plot is an effective tool for measuring codon usage bias [35]. Our ENC-plot analysis showed that mutational pressures can slightly affect the formation of codon usage bias. However, our neutrality plot analysis indicates that natural selection might play an important role in shaping the codon usage bias. This phenomenon also exists in other species, such as *Arabidopsis thaliana*, *Drosophila melanogaster*; and *Caenorhabditis elegans*, in which natural selection is also highly important in shaping codon usage bias in the complete genome [36]. Through previous research we know that *gp64* is a highly conserved gene [37]. It is also one of the homologue of NPV

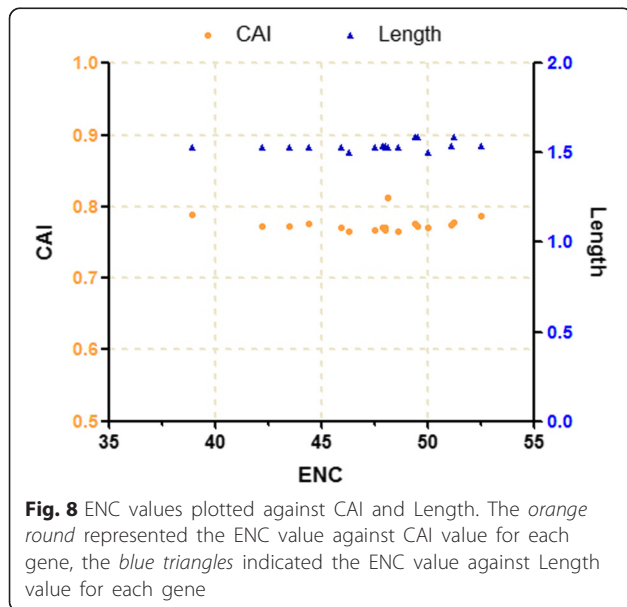


Fig. 8 ENC values plotted against CAI and Length. The orange round represented the ENC value against CAI value for each gene, the blue triangles indicated the ENC value against Length value for each gene

glycoprotein. According to the results of this study, the gene length of 18 glycoprotein is almost the identical size, and their CAI values are maintained at the same level. These results illustrate that the length of the glycoprotein and its expression level would not random variation, in other words, glycoprotein is one of the conservative genes in NPV.

Conclusions

Codon usage patterns were similar between different NPV viral species in same genus. Both Neighbor-joining analysis and clustering analysis were showing the similar conclusion. Multiple factors can affect the synonymous codon usage bias of every organism. Through a series of research and analysis, we can draw the following conclusions: The glycoprotein gene of the NPV genus exhibits a weak codon usage bias. Nucleotide composition, mutation pressure, gene length, and gene expression levels all influence synonymous codon usage bias, with natural selection being the main influence factor. Though codon usage bias is not a necessary metric for carrying out traditional phylogenetic analysis, our study enables us to understand the molecular and genetic mechanisms of viral evolution from a novel perspective. Future advances in the understanding of codon usage evolution will undoubtedly aid us in achieving a more nuanced mastery of viral genetics.

Methods

Codon usage bias measurement index

The effective number of codons (ENC) is a measure that quantifies the extent to which the usage of a gene departs from the equal usage of synonymous codons. It is an excellent indicator of codon usage bias in both genes and genomes. The minimum ENC value is 20, indicating severe codon usage bias, and the maximum value is 61, indicating equally likely usage of all codons.

Relative synonymous codon usage (RSCU) refers to a relative ratio that describes the usage frequency of one specific codon compared to the usage frequency of synonymous codon for the same corresponding amino acid. If the RSCU value is 1, codons are used equally with no bias. Codons with an RSCU value greater than 1 exhibit strong bias (i.e., used more frequently than other synonymous codons), whereas codons with an RSCU value less than 1 exhibit negative bias and are used less frequently than other synonymous codons.

The codon adaptation index (CAI) is another effective measure of codon usage bias, in which each codon is referenced to an optimal codon frequency derived from a set of highly expressed genes. CAI values range from 0 to 1. A value of one indicates strong codon bias in which the optimal codon is always used, and vice versa.

Multifactor variable analysis

Correspondence analysis (COA) is a widely used statistical method used in the analysis of multiple factors and their influences on a particular component. With respect to our experiment, correspondence analysis was used to analyze the effects of various factors on the formation of synonymous codon usage bias in various genes.

Linear regression analysis (LRA) and factor analysis (FA) were used to analyze the relationship between the ENC values and GC3 content of glycoprotein in NPV and the level of correlation between GC12 content and GC3 content. This analysis allowed us to deduce the effects of mutational pressure on codon bias formation.

Neighbor joining (NJ) is a bottom-up clustering method for the creation of phylogenetic trees. Usually used for trees on DNA or protein sequence data, the algorithm knowledge of the distance between each pair of taxa to form the tree.

Cluster analysis (CA) is an analytical method that divides data into groups in such a way that elements more similar to each other are grouped together. Distance is not constant in cluster analysis. The euclidean distance which describes the linear correlation between two variables, was used in our analysis to determine distance.

ENC-plot analysis was used to determine the decisive factors affecting codon usage bias. Each point in the plot corresponds to a GC3s value of a particular gene. Sets of points located on the standard curve indicate mutational pressure determines codon usage bias. Alternatively, points located below the standard curve indicate there are other factors other than mutational pressure affecting codon usage bias.

A neutrality plot analysis was used to determine the extent to which mutational pressures affect codon usage bias as compared to natural selection. Synonymous codon mutations often occur in the third position of the codon, though at times mutations may also occur in the first and second positions, leading to non-synonymous codons. Using GC3 as a horizontal coordinate and GC12 as a vertical coordinate, the GC3 and GC12 contents of glycoprotein genes were plotted and a regression line was calculated to determine the extent to which mutational pressures played a role in the formation of codon usage bias as opposed to natural selection. Regression lines that fall near the diagonal (slope = 1) indicate weak external selection pressures on the generation of codon usage bias, whereas regression curves deviating from the diagonal indicate a heavy influence of natural selection on codon usage bias.

Software

All indices of codon usage bias above were calculated from the data set using the program CodonW 1.4.4 (<http://codonw.sourceforge.net/>). Clustering analysis and

correlations between codon usage variations amongst indices of codon usage were carried out using a statistical software called SPSS Version 22.0, MEGA 6.0, ClustalX 2.0 and GraphPad Prism 5.0.

Additional file

Additional file 1: Table S1. All the indices of total genes. (XLS 27 kb)

Acknowledgments

The authors thank Prof. Christos Louis for critical suggestions of the manuscript and thank Lei Wei and Xian Jia for amendments before submission. We also would like to thank Dr. Bing Han, Vincent Tu (Albert Einstein College of Medicine) and Jonathan Jih (University of California, Los Angeles) for their assistance in preparing this manuscript.

Funding

This research was financially supported by the National Natural Science Foundation of China (No.31372373 to JS), the Natural Science Foundation of Guangdong Province, China (No.2016A030311018 to JS), Science and Technology Planning Project of Guangdong Province (China No.2013B090500118 to JS), Science and Technology Planning Project of Guangzhou (China No.201510010276 to JS).

Availability of data and materials

The sequence set of the 18 glycoprotein genes of different NPV species used in correspondence analysis were obtained from National Center for Biotechnology Information (NCBI, <http://www.ncbi.nlm.nih.gov/>). All of the 18 glycoprotein genes used in clustering analysis/Neighbor joining were achieved from NCBI (accession number: NC_008035.3, NC_023177.1, NC_021925.1, NC_021924.1, NC_003083.1, NC_007767.1, NC_008725.1, NC_001875.2, NC_004323.1, NC_012672.1, NC_001962.1, NC_008520.1, NC_001623.1, NC_005137.2, NC_004778.3, NC_026430.1, NC_008349.1, NC_019945.1) and they were the same as Additional file 1.

Authors' contributions

YZ, HZ and JS conceived the study. YZ and JS carried out the manuscript and writing work. AX gave the Genetic and biological statistical analysis. YZ, HZ, DY, ZJ and QQ performed the data analysis. HZ coordinated the sampling of material. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

¹Subtropical Sericulture and Mulberry Resources Protection and Safety Engineering Research Center, Guangdong Provincial Key Laboratory of Agro-animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou 510642, People's Republic of China. ²Sericultural Research Institute, Chinese Academy of Agricultural Sciences, Zhenjiang Jiangsu 212018, People's Republic of China.

Received: 2 June 2016 Accepted: 16 August 2016

Published online: 24 August 2016

References

- Ingvarsson PK. Molecular evolution of synonymous codon usage in *Populus*. *BMC Evol Biol.* 2008;8:307.
- Lavner Y, Kotlar D. Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene.* 2005;345(1):127–38.

3. Jia J, Xue Q. Codon usage biases of transposable elements and host nuclear genes in *Arabidopsis thaliana* and *Oryza sativa*. *Genomics Proteomics Bioinformatics*. 2009;7(4):175–84.
4. Baeza M, Alcaino J, Barahona S, Sepulveda D, Cifuentes V. Codon usage and codon context bias in *Xanthophyllomyces dendrorhous*. *BMC Genomics*. 2015;16:293.
5. Yang X, Luo XN, Cai XP. Analysis of codon usage pattern in *Taenia saginata* based on a transcriptome dataset. *Parasites Vectors*. 2014;7:527.
6. Duret L. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev*. 2002;12(6):640–9.
7. Duan X, Yi S, Guo X, Wang W. A comprehensive analysis of codon usage patterns in blunt snout bream (*Megalobrama amblycephala*) based on RNA-Seq Data. *Int J Mol Sci*. 2015;16(6):11996–2013.
8. Duret L, Mouchiroud D. Expression pattern, and surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc Natl Acad Sci U S A*. 1999;96(8):4482–7.
9. Tsai CJ, Sauna ZE, Kimchi Sarfaty C, Ambudkar SV, Gottesman MM, Nussinov R. Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *J Mol Biol*. 2008;383(2):281–91.
10. Marin M. Folding at the rhythm of the rare codon beat. *Biotechnol J*. 2008;3(8):1047–57.
11. Xiao JF, Yu J. A scenario on the stepwise evolution of the genetic code. *Genomics Proteomics Bioinformatics*. 2007;5(3-4):143–51.
12. Bulmer M. The selection-mutation-drift theory of synonymous codon usage. *Genetics*. 1991;129(3):897–907.
13. Shah P, Gilchrist MA. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proc Natl Acad Sci U S A*. 2011;108(25):10231–6.
14. Supek F, Skunca N, Repar J, Vlahovicek K, Smuc T. Translational selection is ubiquitous in prokaryotes. *Plos Genetics*. 2010;6(6):e1001004.
15. Palidwor GA, Perkins TJ, Xia X. A general model of codon bias due to GC mutational bias. *Plos One*. 2010;5(10):e13431.
16. Hambuch TM, Parsch J. Patterns of synonymous codon usage in *Drosophila melanogaster* genes with sex-biased expression. *Genetics*. 2005;170(4):1691–700.
17. Hu J, Zhao X, Zhang Z, Yu J. Compositional dynamics of guanine and cytosine content in prokaryotic genomes. *Res Microbiol*. 2007;158(4):363–70.
18. Akashi H. Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. *Gene*. 1997;205(1-2):269–78.
19. Goodarzi H, Torabi N, Najafabadi HS, Archetti M. Amino acid and codon usage profiles: adaptive changes in the frequency of amino acids and codons. *Gene*. 2008;407(1-2):30–41.
20. Berg OG. Selection intensity for codon bias and the effective population size of *Escherichia coli*. *Genetics*. 1996;142(4):1379–82.
21. Prat Y, Fromer M, Linial N, Linial M. Codon usage is associated with the evolutionary age of genes in metazoan genomes. *BMC Evol Biol*. 2009;9:285.
22. Sharp PM, Matassi G. Codon usage and genome evolution. *Curr Opin Genet Dev*. 1994;4(6):851–60.
23. Qin H, Wu WB, Comeron JM, Kreitman M, Li WH. Intra-genetic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics*. 2004;168(4):2245–60.
24. Kadlec J, Loureiro S, Abrescia NGA, Stuart DI, Jones IM. The postfusion structure of baculovirus gp64 supports a unified view of viral fusion machines. *Nat Struct Mol Biol*. 2008;15(10):1024–30.
25. Blissard GW, Wenz JR. Baculovirus gp64 envelope glycoprotein is sufficient to mediate pH-dependent membrane fusion. *J Virol*. 1992;66(11):6829–35.
26. Lang CY, Song JH, Chen XW. The GP64 protein of *Autographa californica* multiple nucleopolyhedrovirus rescues *Helicoverpa armigera* nucleopolyhedrovirus transduction in mammalian cells. *J Gen Virol*. 2005;86:1629–35.
27. Monsma SA, Oomens AGP, Blissard GW. The GP64 envelope fusion protein is an essential baculovirus protein required for cell-to-cell transmission of infection. *J Virol*. 1996;70(7):4607–16.
28. Jiang L, Xia Q. The progress and future of enhancing antiviral capacity by transgenic technology in the silkworm *Bombyx mori*. *Insect Biochem Mol Biol*. 2014;48:1–7.
29. Jiang L, Zhao P, Wang G, Cheng T, Yang Q, Jin S, Lin P, Xiao Y, Sun Q, Xia Q. Comparison of factors that may affect the inhibitory efficacy of transgenic RNAi targeting of baculoviral genes in silkworm, *Bombyx mori*. *Antiviral Res*. 2013;97(3):255–63.
30. Jia X, Liu S, Zheng H, Li B, Qi Q, Wei L, Zhao T, He J, Sun J. Non-uniqueness of factors constraint on the codon usage in *Bombyx mori*. *BMC Genomics*. 2015;16:356.
31. Kawabe A, Miyashita NT. Patterns of codon usage bias in three dicot and four monocot plant species. *Genes Genet Syst*. 2003;78(5):343–52.
32. Peixoto L, Fernandez V, Musto H. The effect of expression levels on codon usage in *Plasmodium falciparum*. *Parasitology*. 2004;128:245–51.
33. Zhao X, Huo KK, Li YY. Synonymous codon usage in *Pichia pastoris*. *Chin J Biotechnol*. 2000;16(3):308–11.
34. Tao P, Dai L, Luo M, Tang F, Tien P, Pan Z. Analysis of synonymous codon usage in classical swine fever virus. *Virus Genes*. 2009;38(1):104–12.
35. Wright F. The 'effective number of codons' used in a gene. *Gene*. 1990;87(1):23–9.
36. Qiu S, Bergero R, Zeng K, Charlesworth D. Patterns of codon usage bias in *Silene latifolia*. *Mol Biol Evol*. 2011;28(1):771–80.
37. Rahman MM, Gopinathan KP. Characterization of the gene encoding the envelope fusion glycoprotein GP64 from *Bombyx mori* nucleopolyhedrovirus. *Virus Res*. 2003;94(1):45–57.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

