Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Theoretical Biology 355 (2014) 128–139

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

System analysis of synonymous codon usage biases in archaeal virus genomes

Sen Li, Jie Yang*

State Key Laboratory of Pharmaceutical Biotechnology, School of Life Science, Nanjing University, Nanjing 210093, China

HIGHLIGHTS

- The SCUB of archaeal virus genes depends mainly on GC richness of genome.
- The mutational pressure is the main factor that influences SCUB.
- The aromaticity of each protein is also critical in affecting SCUB.
- The translational selection could play a leading role in HRPV1's SCUB.
- The mode is helpful to explore the origin of life and the evolution of biology.

ARTICLE INFO

Article history:
Received 29 September 2013
Received in revised form
11 March 2014
Accepted 12 March 2014
Available online 28 March 2014

Keywords:
Mutational bias
Gene function
Hierarchical cluster analysis
Evolution

ABSTRACT

Recent studies of geothermally heated aquatic ecosystems have found widely divergent viruses with unusual morphotypes. Archaeal viruses isolated from these hot habitats usually have double-stranded DNA genomes, linear or circular, and can infect members of the Archaea domain. In this study, the synonymous codon usage bias (SCUB) and dinucleotide composition in the available complete archaeal virus genome sequences have been investigated. It was found that there is a significant variation in SCUB among different Archaeal virus species, which is mainly determined by the base composition. The outcome of correspondence analysis (COA) and Spearman's rank correlation analysis shows that codon usage of selected archaeal virus genes depends mainly on GC richness of genome, and the gene's function, albeit with smaller effects, also contributes to codon usage in this virus. Furthermore, this investigation reveals that aromaticity of each protein is also critical in affecting SCUB of these viral genes although it was less important than that of the mutational bias. Especially, mutational pressure may influence SCUB in SIRV1, SIRV2, ARV1, AFV1, and PhiCh1 viruses, whereas translational selection could play a leading role in HRPV1's SCUB. These conclusions not only can offer an insight into the codon usage biases of archaeal virus and subsequently the possible relationship between archaeal viruses and their host, but also may help in understanding the evolution of archaeal viruses and their gene classification, and more helpful to explore the origin of life and the evolution of biology.

1. Introduction

Universal phylogenetic tree in rooted form showed that life on this planet would comprise three domains, Bacteria, Archaea, and Eucarya (Woese et al., 1990). Viruses can infect members of domains Bacteria, Eukarya, and Archaea. Compared with the bacterial and eukaryal domains, little is known about archaeal domain and their viruses (Snyder et al., 2013). Archaeal domain is composed of two major kingdoms: *Crenarchaeota* and *Euryarchaeota* while the viruses of Archaea are identified prior to appreciation of the existence of domain Archaea itself (Woese et al., 1990). Interestingly, Archaea appear to contain a combination of bacterial and eukaryotic features. The cell structure and metabolic functions of Archaea more closely resemble Bacteria, whereas the information processing in Archaea, such as deoxyribonucleic acid (DNA) replication, transcription, and translation, share more similarities with Eukarya (Snyder et al., 2013).

There are four kinds of nitrogenous bases in DNA: adenine (A), guanine (G), cytosine (C), and thymine (T). The approximate equimolarities $[A] \approx [T]$ and $[G] \approx [C]$ for single stranded DNA molecules had experimentally been determined, which was later

* Correspondence to: Department of Biochemistry, Nanjing University, Nanjing 210093, China. Tel.: +86 25 8359 4060; fax: +86 25 8332 4605.
E-mail address: luckyjyj@sina.com.cn (J. Yang).

http://dx.doi.org/10.1016/j.jtbi.2014.03.022
0022-5193/

proved theoretically that these equalities (referred to as the second parity rule, or PR2) should be observed at the equilibrium state, when mutational and selective pressures are symmetric with respect to the 2 DNA strands (Necşulea and Lobry, 2007). On the other hand, DNA has at least two functions: (i) to provide special sequences for encoding gene products or for regulating transcription and (ii) to provide for genome replication and segregation (Karlin et al., 1994). While the former requires some sequence specificity, the latter may be mostly DNA structure specific. The genome putatively requires compositional flexibility and balance and conveys controls and information in terms of both DNA structure and sequence. GC skews and AT skews are widely encountered in prokaryotic genomes, while the switch in skew direction generally occurs at the origin and terminus of replication (Necşulea and Lobry, 2007). Strand compositional asymmetry may be the consequence of differences in replication synthesis of the leading versus lagging strand, on differences between template and coding strand associated with transcription-coupled repair mechanisms or deamination events, on differences in promoter and gene density between the two strands, on differences in residue and codon biases depending perhaps on gene function, expression level, or operon organization, or on differences in single-base or context-dependent mutational rates (Mrázek and Karlin, 1998). Moreover, whole-genome inverse duplication also contributes to strand asymmetry in bacterial genomes (Sanchez and Jose, 2002; Albrecht-Buehler, 2006; Kong et al., 2009). It has been demonstrated that DNA replication-associated and transcription-associated mutation bias and/or selective codon usage bias (CUB) may affect the strand nucleotide composition asymmetrically in eukaryotic genomes (Niu et al., 2003).

All amino acids except Met and Trp are coded by more than one codon within the standard genetic codes, which are called synonymous codons that are not used randomly in intergenome (Grantham et al., 1980; Nakamura et al., 1991; Angellotti et al., 2007). The frequencies of individual synonymous codons are quite variable from genome to genome and within genomes, from gene to gene (Kurland, 1991). Synonymous codon usage, studied in a number of living organisms, has basically proven to be non-random and species-specific. Several factors, including directional mutational bias, translational selection, secondary protein structure, replicational and transcriptional selection, and environmental factors, have been shown to influence codon usage in a variety of organisms (Sau et al., 2007). Synonymous codon usage biases (SCUB) are associated with various biological factors, such as gene expression level, gene length, gene translation initiation signal, protein amino acid composition, protein structure, tRNA abundance, mutation frequency and patterns, and GC compositions (Angellotti et al., 2007). Mutational biases influence the genome base composition, which in turn affects codon usage considerably (Palidwor et al., 2010). Asymmetry in DNA replication and repair of the leading and lagging DNA strands (Sueoka, 1962) creates further codon usage bias. Furthermore, it has been shown that certain types of SCUB are associated with gene expression levels (Holm, 1986). This association is often explained through selection towards translational efficiency (Lavner and Kotlar, 2005), which regards optimization of both translation rate (Varenne et al., 1984) and fidelity (Precup and Parker, 1987). Selection towards codon usage that promotes efficient translation has both local effects on specific genes (Drummond and Wilke, 2009) as well as global effects on the organism's fitness. The latter is achieved by promoting rapid recycling of ribosomes, reducing costs wasted on correcting translation errors (Ibba and Soll, 1999) and lowering the production of inactive and at times toxic proteins (Zaher and Green, 2009; Wald et al., 2012).

Additionally, gene function (Ma et al., 2002a) and protein secondary structure (Ma et al., 2002b; Kahali et al., 2007) can also be related to codon usage. Quantification of CUB, especially at genomic scale, helps understand evolution of living organisms (Angellotti et al., 2007). Investigation of codon usage patterns and causes of CUB can provide a basis for understanding the viral molecular evolution, particularly the interaction between viruses and their host (Shackelton et al., 2006). After extensive studies in CUB and nucleotide composition of bacterial, yeast, fruit fly and mammals, several researchers focused on codon usage in virus (Mooers and Holmes, 2000). For instance, it has been observed that codon usage preference is related to mutational pressure, G+C content, the route of transmission of the virus and the segmented nature of the genome in human RNA viruses (Jenkins and Holmes, 2003). As to vertebrate DNA viruses, genome-wide mutational pressure, rather than natural selection is the main factor to determine codon usage (Shackelton et al., 2006). The correlation between codon usage and tRNA availability is found in analysis of the bovine papillomavirus type 1(BPV1) late genes (Zhou et al., 1999). It has been reported that codon preference in host genome can also strongly influence replication and gene expression of its corresponding virus (Zhao et al., 2003). The relationship of genotypes, virulence, dinucleotides and codon usage has been investigated in classical swine fever viruses (CSFV) (Pan et al., 2009).

Hyperthermophiles that thrive at temperatures greater than 80 °C, halophiles and anaerobic methanogens are the main life-styles within the domain Archaea. The domain Archea is infected by viruses (Prangishvili et al., 2006) and the archaeal viruses represent approximately 1% of over 5500 prokaryotic viruses that are currently known (Ackermann, 2007). Sequence similarities between genes of archaeal viruses are generally very limited, and moreover there are no other matches of most predicted genes in public sequence databases (Prangishvili and Garrett, 2004). Almost all isolated archaeal viruses contain a double-stranded DNA (dsDNA) genome with one exception: *Halorubrum* pleomorphlc virus 1 (HRPV-1) which has been found recently with a single-stranded DNA (ssDNA) (Pietilä et al., 2009). In archaeal virus with dsDNA genomes, some codons exist exclusively on one DNA stand while other codons on both DNA strands (Haring et al., 2004). There is a growing interest in Archaea and their viruses because their habitats elucidate the limits where life is possible.

Although genome sequences of several archaeal virus species have been published and many studies have been performed on it in recent years (Tolstrup et al., 2000), no in depth genome analyses have so far been made on codon usage, which may provide more information about these unique creatures. In this investigation, we have analyzed and compared the codon usage data of 11 available complete genome sequences of archaeal virus. Such information not only can offer an insight into the codon usage pattern of archaeal virus and subsequently the possible relationship between archaeal virus and its host, but also is more helpful to explore the origin of life and the evolution of biology.

2. Materials and methods

11 available complete genome sequences of archaeal virus were downloaded from NCBI (⟨http://www.ncbi.nlm.nih.gov/⟩), involving sulfolobus virus Kamchatka 1 (SSVK1), sulfolobus virus 2 (SSV2), sulfolobus virus 1 (SSV1), sulfolobus spindle-shaped virus Ragged Hills (SSVRH), sulfolobus islandicus rod-shaped virus 1 (SIRV1), sulfolobus islandicus rod-shaped virus 2 (SIRV2), sulfolobus spindle-shaped virus 4 (SSSV4), acidianus rod-shaped virus 1 (ARV1), acidianus filamentous virus 1 (AFV1), halorubrum pleomorphic virus 1 (HRPV1), and natrialba phage PhiCh1 (PhiCh1) (Table 1). The former nine belongs to viruses in the kingdom Crenarchaeota, while the latter two pertains to viruses in the

Table 1
Eleven complete genome sequences of Archaeal virus under study.

GSN	Strain	Acronym	DNA	Host [a] (genus [domain])	Accession no.	Group	Genome (bp)	GC content (%)	Ref.
I	Sulfolobus (spindle-shaped) virus Kamchatka 1	SSVK1	dsDNA	Sulfolobus [Crenarchaeota]	NC_005361	1	17,385	38	Wiedenheft et al. (2004)
II	Sulfolobus (spindle-shaped) virus 2	SSV2	dsDNA	Sulfolobus [Crenarchaeota]	NC_005265	1	14,796	38	Stedman et al. (2003)
III	Sulfolobus (spindle-shaped) virus 1	SSV1	dsDNA	Sulfolobus [Crenarchaeota]	NC_001338	1	15,465	39	Palm et al. (1991)
IV	Sulfolobus (spindle-shaped) virus Ragged Hills	SSVRH	dsDNA	Sulfolobus [Crenarchaeota]	NC_005360	1	16,473	37	Wiedenheft et al. (2004)
V	Sulfolobus islandicus rod-shaped virus 1	SIRV1	dsDNA	Sulfolobus [Crenarchaeota]	NC_004087	2	32,308	25	Peng et al. (2001)
VI	Sulfolobus islandicus rod-shaped virus 2	SIRV2	dsDNA	Sulfolobus [Crenarchaeota]	NC_004086	2	35,450	25	Peng et al. (2001)
VII	Sulfolobus spindle-shaped virus 4	SSSV4	dsDNA	Sulfolobus [Crenarchaeota]	NC_009986	1	15,135	38	Peng (2008)
VIII	Acidianus rod-shaped virus 1	ARV1	dsDNA	Acidianus [Crenarchaeota]	NC_009965	1	24,655	39	Vestergaard et al. (2005)
IX	Acidianus filamentous virus 1	AFV1	dsDNA	Acidianus [Crenarchaeota]	NC_005830	1	20,869	36	Bettstetter et al. (2003)
X	Halorubrum pleomorphic virus 1	HRPV1	ssDNA	Halorubrum [Euryarchaeota]	NC_012558	3	7048	54	Pietilä et al. (2009)
XI	Natrialba phage PhiCh1	PhiCh1	dsDNA	Natrialba [Euryarchaeota]	NC_004084	3	58,498	61	Klein et al. (2002)

Note: GSN: genome serial number.

[a] Host lineage: Archaea, Crenarchaeota, Thermoprotei, Sulfolobales, Sulfolobaceae, Sulfolobus.
Archaea, Crenarchaeota, Thermoprotei, Sulfolobales, Sulfolobaceae, Acidianus.
Archaea, Euryarchaeota, Halobacteria, Halobacteriales, Halobacteriaceae, Halorubrum.
Archaea, Euryarchaeota, Halobacteria, Halobacteriales, Halobacteriaceae, Natrialba.

kingdom Euryarchaeota. The genome serial number (GSN), strain, their responding host, Genbank accession numbers, genome length, G+C content and references are listed in Table 1. From these genomes, we only extract the predicted ORFs that show significant matches to sequences in public databases. In addition, to minimize the sampling error, genes either with $>99\%$ sequence identities or have internal termination codons were excluded. Also, we only retain those genes, which are greater than or equal to 300 bp. Finally 26 genes were selected for analysis (Table 2).

Correlation analysis was carried out using the Spearman's rank correlation analysis method. In order to compare the variation of codon usage among different gene groups, one-way analysis of variance (one-way ANOVA) and multiple comparison have been used. Using a hierarchical cluster method the cluster analysis and the distances between selected sequences were calculated by the Euclidean distance method. All statistical analyses, as well as cluster analysis, were implemented using the statistical analysis software SPSS Version 11.5.

2.1. Measures of synonymous codon usage bias

To normalize codon usage within data sets of differing amino acid compositions, relative synonymous codon usage (RSCU) values, which are particularly useful in comparing codon usage between genes, were calculated by dividing the observed codon usage by that expected when all codons for the same amino acid are used equally (Paul and Wen-Hsiung, 1986). The "effective number of codons" (ENC) was often used to measure the non-uniformity of synonymous codon usage, which generates values ranging from 20 for a gene with extreme bias that only one synonymous codon is used for each amino acid, to 61 for a gene with no bias using synonymous codons equally (Wright, 1990). The index GC3S was used to analyze the extent of base composition bias by calculating the fraction of nucleotides G+C at the synonymous third codon position (excluding start codons, Trp and the termination codons). In addition, the frequency of aromatic amino acids (Aromo) in the hypothetical genes was also computed. All the indices mentioned above were calculated using the program CodonW version 1.4, available from ⟨http://www.molbiol.ox.ac.uk/cu⟩.

2.2. Correspondence analysis (COA)

As implemented in CodonW, Correspondence analysis was used to explore the variation of RSCU values among Archaeal virus genes. In this multivariate statistical analysis, all genes were plotted in a 59-dimensional hyperspace according to their usage of the 59 sense codons (excluding start codons, Trp and stop codons). Major trends within this data set can be identified using measures of relative inertia and genes ordered according to their positions along the axis of major inertia.

2.3. Relative dinucleotide abundance in archaeal virus ORFs

By using the method described by Karlin and Burge (1995), the relative abundance of dinucleotides in archaeal virus ORFs was assessed. The odds ratio $\rho_{xy}=f_{xy}/f_x f_y$, where f_x denotes the frequency of the nucleotide X and f_{xy} the frequency of the dinucleotide XY, etc., for each dinucleotide were calculated. If $\rho_{xy}>1.23$ (or <0.78), the XY pair is considered to be of high (or low) relative abundance compared with a random association of mononucleotides.

3. Results

3.1. Synonymous codon usages in archaeal virus

The details of genes in different archaeal virus species and the overall RSCU values of 59 codons were, respectively, shown in Tables 2 and 3. According the G+C content, 11 archaeal viruses are divided into 3 groups, which is indicated in Tables 1 and 2. In group 2 with genome G+C content of 25%, most of the preferentially used codons are A- or U-ended codons. On the other hand, C- or G-ended

Table 2
Selected genes which have certain matches in public sequence database.

SN	Group	GSN	ENC	GC3s[a]	GC[b]	F1[c]	F2[d]	Aromo[e]	GA	SN	Group	GSN	ENC	GC3s	GC	F1	F2	Aromo	GA
1	1	I	58.73	0.361	0.367	0.246892	0.178836	0.17316	f	**14**	2	VI	37.62	0.183	0.358	0.64488	−0.24913	0.104478	i
2	1	I	47.89	0.439	0.408	0.192363	0.360043	0.088235	g	**15**	1	VII	58.31	0.434	0.377	−0.02827	0.083677	0.065421	h
3	1	I	61	0.446	0.415	0.058542	0.478179	0.065041	h	**16**	1	VII	55.91	0.457	0.413	0.073852	0.219034	0.137339	f
4	1	II	57.57	0.489	0.415	0.100366	0.181638	0.141631	f	**17**	1	VII	59.74	0.447	0.416	0.127912	0.191922	0.087379	g
5	1	II	54.22	0.405	0.403	0.303551	0.142189	0.097561	g	**18**	1	VIII	61	0.445	0.422	0.009764	−0.04844	0.10628	g
6	1	II	60	0.459	0.393	−0.01397	0.279834	0.071429	h	**19**	1	IX	38.42	0.327	0.381	0.372225	−0.27826	0.107623	g
7	1	III	52.09	0.39	0.378	0.364101	0.206763	0.111554	f	**20**	3	X	46.01	0.546	0.548	−0.32849	−0.07058	0.046647	k
8	1	IV	46.01	0.329	0.368	0.422873	0.128048	0.125506	f	**21**	3	XI	35.83	0.86	0.63	−0.96386	−0.11877	0.08137	l
9	1	IV	52.93	0.399	0.389	0.301396	0.147137	0.097561	g	**22**	3	XI	37.08	0.833	0.625	−0.85322	−0.16631	0.058594	m
10	2	V	36.11	0.083	0.214	0.740341	−0.30574	0.149758	g	**23**	3	XI	33.7	0.883	0.625	−1.01563	−0.2054	0.048583	n
11	2	V	45.39	0.229	0.376	0.592535	0.071706	0.104478	i	**24**	3	XI	38.28	0.857	0.641	−0.85399	−0.11729	0.081272	o
12	2	V	32.45	0.15	0.264	0.645066	−0.56159	0.132231	j	**25**	3	XI	40.96	0.784	0.629	−0.79422	−0.10627	0.082589	p
13	2	VI	31.77	0.084	0.219	0.754924	−0.35846	0.149758	g	**26**	3	XI	34.71	0.86	0.627	−1.00412	−0.00395	0.081818	q

Note: SN: sequence number; GSN: genome serial number; ENC: effective number of codon; GA: Gene Bank annotation.
[f]Distant but significant similarity to bacterial DnaA, a multifunctional DNA binding protein (replication initiation, transcription regulation).
[g]Putative RecB family exonuclease usually found in association with Clustered regularly interspaced short palindromic repeats (CRISP).
[h]Putative helix-turn-helix transcription protein. Part of the early transcription unit.
[i]Major structural protein. Based on virion structure and high isoelectric point, these proteins likely interact directly with DNA.
[j]Similar to archaeal holliday junction resolvases.
[k]Putative ATPase.
[l]Putative portal protein.
[m]Contains ATP-binding motif.
[n]Putative proliferation cellular nuclear antigene.
[o]Putative C5-cytosine methyltransferase.
[p]Putative N4-cytosine methyltransferase.
[q]Putative three transmembrane helices protein.

[a] The frequency of G+C at the third synonymously variable coding position.
[b] The frequency of G+C of this gene.
[c] The first axis values of each gene in COA.
[d] The second axis values of each gene in COA.
[e] The aromaticity value of each protein.

Table 3
Number of codons with RSCU > 1 in for each amino acid in different groups.

		Phe	Leu	Ile	Val	Ser	Pro	Thr	Ala	Tyr	His	Gln	Asn	Lys	Asp	Glu	Cys	Arg	Gly
Group 1	A- or U-ended	8	17	12	16	25	12	21	17	11	7	6	4	8	11	10	4	17	12
	C- or G-ended	5	18	4	6	9	10	6	6	2	1	4	10	4	2	2	0	15	13
Group 2	A- or U-ended	5	10	5	8	9	9	9	8	5	3	4	5	5	5	5	3	5	9
	C- or G-ended	0	2	0	1	3	0	0	1	0	2	0	0	0	0	0	0	1	1
Group 3	A- or U-ended	0	0	1	2	2	1	1	1	0	0	0	0	2	0	0	1	4	1
	C- or G-ended	7	13	7	7	17	12	11	13	7	7	6	7	5	7	7	6	13	9

codons are used with higher frequency in the genes from group 3 which harbors two GC rich genomes. As to group 1 with G+C content ranging from 36% to 39%, it seems that A- or U-ended codons are more popular than C- or G-ended codons although this trend is not absolute. In group 3 with GC-rich genome, C- or G-ended codons are used with higher frequency. The above phenomenon is independently summarized in Table 3. As expected, different gene groups have their own A- or U-ended and C- or G-ended codons preference due to compositional constraints. Similar results were obtained by analyzing all valid 320 archaeal virus genes from the 11 genomes as described in Supplementary Table 3. Moreover, in the selected 26 archaeal virus genes, there are 6 amino acids preferring codons ended with U in SIRV1 and SIRV2 of group 2, including Thr (ACU), Phe (UUU), Tyr (UAU), Asn (AAU), Asp (GAU), and Cys (UGU), as well as 4 amino acids preferring codons ended with A, such as Pro (CCA), Lys (AAA), Glu (GAA), and Arg (AGA). Particularly, 7 amino acids prefer codons ended with C in PhiCh1 and HRPV1 of group 3, including Asn (AAC), Asp (GAC), His (CAC), Tyr (UAC), Phe (UUC), Cys (UGC), Ile (AUC), and Leu (CUC), whereas 2 amino acids prefer codons ended with G, such as Gln (CAG) and Glu (GAG), besides Met (AUG) and Trp (UGG). About group 1, 3 amino acids mostly choose codons ended with U, including His (CAU), Tyr (UAU), and Asp (GAU), whereas 3 amino acids mainly choose codons ended with A, including Glu (GAA), Lys (AAA), and Leu (UUA). Similarly, different gene groups have their own A- or U-ended and C- or G-ended codons preference due to compositional constraints. The compositional constraints (or mutational bias) and gene functions are the cause of the bias and the effect of natural selection is slight. Categorizing the genomes as well as genes according to their G+C content will facilitate the analysis of SCUB, especially when compositional bias plays a key role in SCUB. We first validated this method by using two groups of known phages with different G+C content: GC-rich Mycobacterium phage and AT-rich Staphylococcus phage, as described in Supplementary Table 4. SCUB analysis indicated that all the genes of phages with similar G+C content were grouped together in various plots (Supplementary Fig. 1A–C). Interestingly, in the dendroid chart generated by using RSCU of each of the selected 104 genes of Mycobacterium and Staphylococcus phages, all these genes were clearly clustered into the high G+C content group (Mycobacterium phage) and low G+C content group (Staphylococcus phage) (Supplementary Fig. 1D), which is consistent with the fact that compositional bias is the major factor

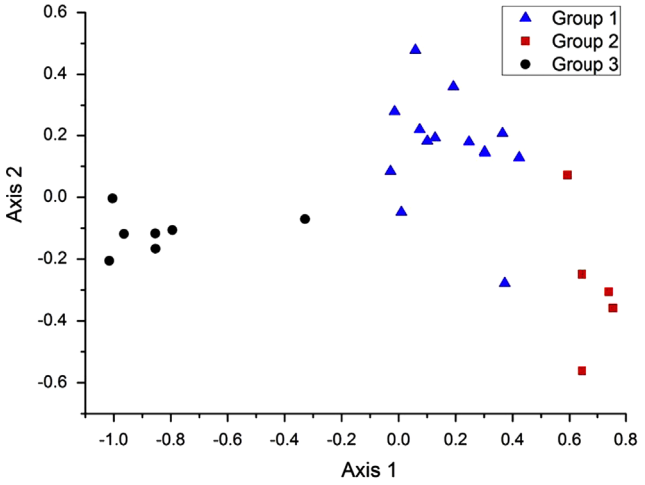

Fig. 1. A plot of value of the first and second axis in COA. The first axis accounts for 48.81% of all variation among ORFs, which is much bigger than other axes (8.76%, 7.16% and 5.62%).

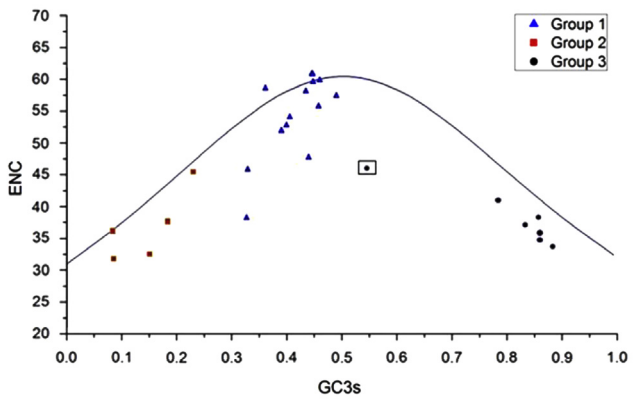

Fig. 2. Effective number of codons used in each ORF plotted against the GC3s. The curve represents the relationship between GC3s and ENC in the absence of selection. The box indicates gene 20 in genome X.

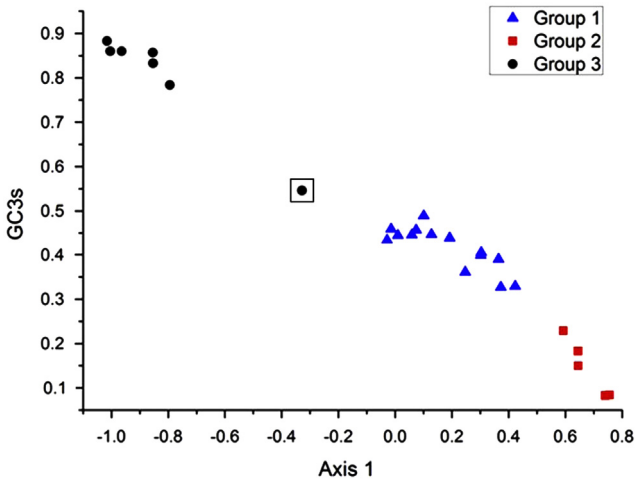

Fig. 3. Effective number of codons used in each ORF plotted against the axis1 values in COA. The box indicates gene 20 in genome X.

influencing SCUB in both Mycobacterium and Staphylococcus phages (Hassan et al., 2009; Sau et al., 2005).

To study the codon usage variation among different archaeal virus genes, ENC and GC3S values of different archaeal virus genes were calculated (Table 2). ENC values of different archaeal virus genes vary from 31.77 to 61, with a mean value of 46.68 and S.D. of 10.36. The data suggests a high heterogeneity of synonymous codon usage among archaeal virus genes selected. This hypothesis is further supported by the GC3s values for each examined archaeal virus genes, which vary from 8% to 88% with a mean of 47% and S.D. of 0.24. In fact, the average GC3s of three gene groups have significant differences as showed by one-way ANOVA and multiple comparison (both at significant level of 0.01). Multiple comparison also indicated that the average ENC value of group 1 differs from that of group 2 and group 3 respectively while there is no difference between the average ENC value of group 2 and group 3. This is probably because genes in group 2 and group 3 have a high codon usage bias with low ENC values, but genes from these two groups tend to use different codons as discussed above.

3.2. Correspondence analysis on codon usages

To study the variation of RSCU values among genes, correspondence analysis (COA) was carried out on these 26 archaeal virus genes examined as a single dataset based on the RSCU value of each gene. Fig. 1 describes the position of each ORF on the plane defined by the first and second principle axes. The first principal axis accounts for 48.81% of the total variation. The next three axes account for 8.76%, 7.16% and 5.62% of the variation respectively. This observation indicates that the first major axis explains much of the variation in trends in codon usage than other axes. It is worth noting that genes separate distinctly according to their group, which suggests that the genome G+C content can have a great influence on codon usage bias in archaeal virus.

3.3. Effect of mutational bias on the codon usage variation in archaeal virus

To study whether the evolution of SCUB is controlled by mutation pressure or natural selection, actual ENC value of each gene was plotted against its GC3s value (Fig. 2). The expected ENC curve shows the situation that codon usage bias is solely due to biased base composition (G+C content). Result showed that points of group 2 mainly lie on GC-poor regions (GC value 0.083 to 0.229) while points of group 3 like to harbor GC-rich regions. The point inside a box represents the only selected gene in genome X with 54% G+C content which is just between group 1 and group 3. Most of the points of ENC values are on or just below the expected curve, which means that genomic GC composition has a profound effect on SCUB although other factors may also affect it. Similar results were found by analyzing all valid 320 archaeal virus genes from the 11 genomes (Supplementary Fig. 2). Moreover, the GC3s values of each gene are plotted against first axis values in COA (Fig. 3). The patterns of codon usage in different genes also seem to be closely related to GC content on the third codon position. Furthermore, this concept is verified by correlation analysis, which has been implemented to find correlation between synonymous codon usage and nucleotide compositions. It is found that coordinates of axis1, the most important axis accounting for the variation, are negative correlated with GC3s and GC respectively ($r=-0.959$, $P<0.01$; $r=-0.899$, $P<0.01$). Taken together, these analyses reveal that most of the SCUB among the selected archaeal virus genes is directly related to the base composition. So, mutational bias may be the major factor accounting for synonymous codon usage variation among genes in these virus genomes.

Table 4
Relative abundance of the 16 dinucleotides in three gene groups.

		Relative abundance of the 16 dinucleotides							
		TT	**TC**	**TA**	**TG**	**CT**	**CC**	**CA**	**CG**
Group 1	Range[a]	1.316 – 0.891	1.25 – 0.541	1.044 – 0.579	1.306 – 0.84	1.595 – 0.753	1.707 – 0.308	1.415 – 0.643	1.344 – 0.348
	Mean ± S.D[b]	1.125 ± 0.124	0.912 ± 0.165	0.91 ± 0.116	1.053 ± 0.136	1.169 ± 0.231	0.859 ± 0.378	1.014 ± 0.218	0.926 ± 0.244
Group 2	Range	1.165 – 1.064	1.441 – 0.652	1.094 – 0.64	1.481 – 0.979	1.12 – 0.824	1.292 – 0.867	1.395 – 1.103	0.715 – 0.154
	Mean ± S.D	1.124 ± 0.041	1.123 ± 0.351	0.842 ± 0.17	1.203 ± 0.187	0.923 ± 0.114	1.119 ± 0.192	1.262 ± 0.124	0.396 ± 0.212
Group 3	Range	1.033 – 0.545	1.905 – 1.397	0.569 – 0.15	1.259 – 0.636	1.438 – 0.912	0.854 – 0.679	1.002 – 0.569	1.439 – 1.152
	Mean ± S.D	0.82 ± 0.196	1.678 ± 0.187	0.351 ± 0.161	0.891 ± 0.206	1.082 ± 0.182	0.765 ± 0.059	0.796 ± 0.147	1.323 ± 0.103
		AT	**AC**	**AA**	**AG**	**GT**	**GC**	**GA**	**GG**
Group 1	Range	1.192 – 0.797	1.369 – 0.625	1.213 – 0.895	1.149 – 0.784	1.087 – 0.677	1.507 – 0.81	1.196 – 0.879	1.261 – 0.854
	Mean ± S.D	0.939 ± 0.103	1.093 ± 0.181	1.019 ± 0.092	0.963 ± 0.111	0.829 ± 0.13	1.091 ± 0.205	1.048 ± 0.086	1.051 ± 0.096
Group 2	Range	1.032 – 0.825	0.869 – 0.607	1.192 – 1.021	1.1 – 0.894	1.095 – 0.732	1.574 – 0.798	1.258 – 0.616	1.142 – 0.739
	Mean ± S.D	0.953 ± 0.078	0.729 ± 0.119	1.081 ± 0.066	1.04 ± 0.085	0.929 ± 0.171	1.293 ± 0.342	0.952 ± 0.304	0.929 ± 0.181
Group 3	Range	1.822 – 0.861	1.13 – 1.02	1.416 – 0.441	1.031 – 0.589	1.045 – 0.834	0.896 – 0.763	1.814 – 1.457	0.894 – 0.656
	Mean ± S.D	1.17 ± 0.338	1.079 ± 0.044	0.878 ± 0.294	0.871 ± 0.157	0.938 ± 0.073	0.82 ± 0.046	1.619 ± 0.107	0.798 ± 0.08

[a] The range of three gene groups' relative dinucleotide ratios.
[b] Mean values of three gene groups' relative dinucleotide ratios ± S.D.

3.4. The relative abundance of dinucleotide also shape the codon usage in archaeal virus

It has been widely reported that dinucleotide bias can affect codon usage (Karlin and Burge, 1995; Zhang et al., 2011; Liu et al., 2010). To investigate the possible effect on the composition of dinucleotide on codon usage in archaeal virus, the relative abundances of the 16 dinucleotides in the 26 archaeal virus genes were calculated. Due to the variation of genomic GC content, these 26 genes are divided into 3 groups as we talked above. As shown in Table 4, although dinucleotides were not randomly distributed and no dinucleotides were present at the expected frequencies, most dinucleotide odds ratios were between 0.78 and 1.23. The relative abundance of CpG in group 2 and UpA, UpC and GpA in group 3 showed the remarkable deviation from the "normal range" (mean ± S.D.= 0.396 ± 0.212, 0.351 ± 0.161, 1.678 ± 0.187 and 1.619 ± 0.107 respectively). The relative abundance of CpA, ApC, and CpG in group 2 and CpC, CpG in group 3 also shows slight deviation from the "normal range" (mean ± S.D.= 1.262 ± 0.124, 0.729 ± 0.119, 1.293 ± 0.342, 0.765 ± 0.059 and 1.323 ± 0.103). Among the 16 dinucleotides 11 were significantly correlated with the first axis value in COA. Among the 5 dinucleotides that are not correlated with the first axis value, three of them are correlated with the second axis value in COA (Table 5). In addition to total dinucleotides, the relative abundances of the 16 intercodon dinucleotides were also calculated (Supplementary Table 5) (Sanchez, 2011). Similarly, significant correlations with the first axis value were observed in 11 out of 16 intercodon dinucleotides. For the 5 intercodon dinucleotides that were not correlated with the first axis, two of them were correlated with the second axis value (Table 5). These observations reveal that the composition of dinucleotides, which are independent of the overall base composition but still the result of differential mutational pressure, also shapes the bias of synonymous codon usage among different archaeal virus ORFs. To further study the possible effect of dinucleotides on codon usage bias, the RSCU value of specific codons, which contain certain dinucleotide, were analyzed. In group 2, among the RSCU value of eight codons that contain the under-represented CpG (CCG, GCG, UCG, ACG, CGC, CGG, CGU and CGA), all (CCG (mean 0.114), GCG (mean 0.366), UCG (mean 0.172), ACG (mean 0.160), CGC (mean 0), CGG (mean 0), CGU (mean 0) and CGA (mean 0)) were markedly suppressed. As to group 3, the RSCU values of twenty four codons that contain UpA (UUA, CUA, AUA, GUA, UAU, UAC, UAA and UAG), UpC (UUC, CUC, AUC, GUC, UCU, UCC, UCA and UCG), and GpA (GAU, GAC, GAA, GAG, UGA, CGA, AGA and GGA) respectively were analyzed. Of these codons, seven codons (UUA (mean 0.017), CUA (mean 0.170), AUA (mean 0.093), GUA (mean 0.171), UAU (mean 0.156), UAA (mean 0) and UAG (mean 0.429)) were suppressed due to the under-representation of UpA. Six codons (UUC (mean 1.901), CUC (mean 3.489), AUC (mean 2.630), GUC (mean 2.563), UCC (mean 1.797) and UCG (mean 1.643)) were over-used because of the over-representation of UpC. Only three codons (GAC (mean 1.537), GAG (mean 1.544), UGA (mean 2.571)) were over-used due to GpA over-representation.

3.5. Cluster analysis

Beyond the analyses mentioned above, a cluster tree was also generated by the hierarchical clustering method based on the variation in RSCU values among 26 archaeal virus genes. As shown in Fig. 4, these 26 genes were divided into 3 distinct sublineages, each of which contains all the genes, which have already been put together as a group. That is to say, genomic GC content, again, was a dominant factor that affects the classification of 26 archaeal virus genes. It should be noted that the distance between gene 20 and other genes in group 3 is farther than that between every two of 21–26 genes. This may be because gene 20 is the only selected gene in genome X with 54% genomic GC content which differs from that of genome XI. Moreover, in a cluster tree generated by RSCU values of all the valid 320 archaeal virus genes from 11 genomes, it was clearly seen that genes belonged to one group tended to cluster together (Supplementary Fig. 3).

3.6. Effect of other factors on codon usage

Although codon bias in archaeal virus can be mainly explained by mutational pressure, there are other factors, with less of an effect, which may also influence the codon usage. To test if any selection pressure contributes to the codon variation between these archaeal virus genes, a correlation analysis was carried out between axis values in COA and aromaticity or GRAVY score of each protein. It was found that axis1 is significantly correlated to aromaticity score ($r = 0.753$, $P < 0.01$), indicating that the frequency of aromatic amino acids (Phe, Tyr and Trp) in the hypothetical translated gene product is also related to the observed variation in codon bias. By using Spearman's correlation, no significant relationship was found between axis values in COA and GRAVY.

From the cluster tree generated above, we can see that genes with similar functions may also display similar SCUB. For an

Table 5
Summary of correlation analysis between the first two axes in COA and sixteen dinucleotides and intercodon dinucleotides in the examined viruses.

		TT	TC	TA	TG	CT	CC	CA	CG	AT	AC	AA	AG	GT	GC	GA	GG
Axis1	r	0.543**	−0.607**	0.586**	0.474*	−0.323	0.621**	0.602**	−0.824**	−0.113	−0.32	0.391*	0.459*	−0.049	0.451*	−0.566**	0.307
	P	0.004	0.001	0.002	0.014	0.107	0.001	0.001	< 0.001	0.582	0.111	0.048	0.018	0.813	0.021	0.003	0.127
Axis2	r	0.238	−0.48*	0.587**	−0.232	0.551**	−0.331	−0.352	0.179	−0.249	0.356	0.124	−0.073	−0.486*	0.24	−0.28	0.571**
	P	0.241	0.013	0.002	0.254	0.004	0.099	0.077	0.38	0.221	0.074	0.546	0.724	0.012	0.238	0.166	0.002
		T/T	T/C	T/A	T/G	C/T	C/C	C/A	C/G	A/T	A/C	A/A	A/G	G/T	G/C	G/A	G/G
Axis1	r	0.683**	0.322	0.636**	0.711**	−0.538**	−0.094	−0.581**	−0.739**	0.805**	0.329	0.578**	0.851**	−0.544**	−0.013	−0.751**	−0.324
	P	< 0.001	0.109	< 0.001	< 0.001	0.005	0.648	0.002	< 0.001	< 0.001	0.101	0.002	< 0.001	0.004	0.951	< 0.001	0.106
Axis2	r	0.124	−0.293	0.488*	−0.147	0.484*	0.192	0.082	0.011	−0.144	0.454*	0.079	−0.183	0.199	0.238	0.321	0.701**
	P	0.546	0.146	0.011	0.473	0.012	0.348	0.689	0.959	0.483	0.020	0.701	0.371	0.329	0.241	0.110	< 0.001

Note: The vertical lines indicate the boundary between codons.
* *P*-value ≤ 0.05.
** *P*-value ≤ 0.01.

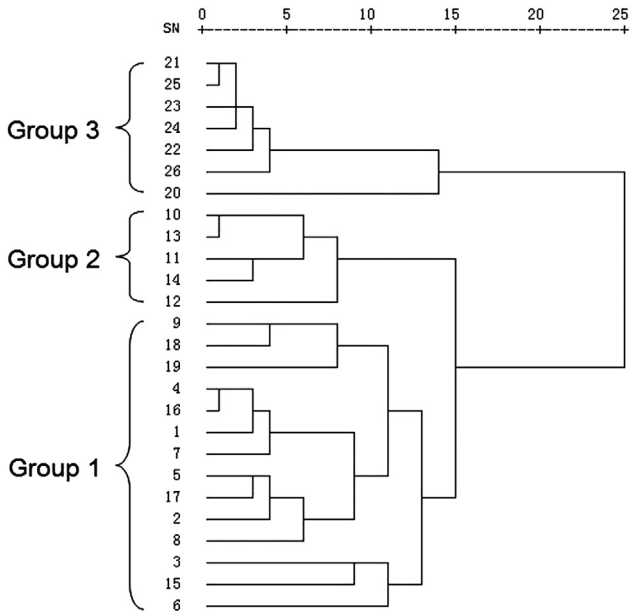

Fig. 4. Dendroid chart of the cluster result of the 26 archaeal virus genes under study based on the hierarchical cluster method.

example, within group 2, genes 10 and 13 tend to cluster together due to their same exonuclease function. Distances between genes 11 and 14 are relatively closer because both of them are major structural protein. On the other hand, distance of gene 12 is quite far from other genes in group 2, which may be because there is no other protein with similar function as Holliday junction resolvase. So, it seems that the distance between genes of similar functions are relatively closer than distances between genes of different functions. But this concept is not absolute because other factors, such as aromaticity, can also determine SCUB at this level.

4. Discussion

SCUB are associated with various biological factors, such as gene expression level, gene length, gene translation initiation signal, protein amino acid composition, protein structure, tRNA abundance, mutation frequency and patterns, and GC compositions (Angellotti et al., 2007). Satapathy et al. (2012) have indicated that tRNA gene numbers might not be the sole determining factor for translational selection of SCUB in bacterial genomes. The patterns of codon usage vary remarkably among organisms, and also among genes from the same genome (Grocock and Sharp, 2002). Codon frequencies can also vary due to mutational biases as well as because of selection (Ran and Higgs, 2012). DNA sequence data from diverse organisms clearly show that synonymous codons for any amino acid are not used with equal frequency, and these biases are a consequence of natural selection during evolution (Angellotti et al., 2007). Main factors that account for these variations were thought to be mutational pressure and translational selection (Shackelton et al., 2006; Karline and Marazek, 1996).

First, mutational pressure in double-stranded DNA genome is the situation when AT to GC substitution rates are not equal to GC to AT substitution rates (Sueoka, 1998). Most substitutions in third codon positions are synonymous, so single nucleotide mutations occurring due to mutational pressure may be fixed in these codon positions by the random genetic drift without any selective limitations. When the level of GC content in third codon positions

(3GC) for most of coding regions in the prokaryotic genome is higher than 0.5, one can suspect that this genome is under the influence of GC pressure (Khrustalev and Barkovsky, 2010). There should be AT pressure in the genome if 3GC level for most of its coding regions is lower than 0.5. The strongest evidence for GC pressure in genome is the situation when 3GC is higher than 1GC and 2GC for most of its genes. If 3GC level is lower than 2GC and 1GC for most of coding regions, there is AT pressure in this genome. Mutational pressure is caused by the imbalance of mutational processes and reparation. The most common and well-studied mutational processes that contribute into mutational pressure are (Sueoka, 2002; Gros et al., 2002): (i) deamination of cytosine leading to C to U transitions, (ii) deamination of methyl-cytosine leading to 5-methyl-C to T transitions (Yoon et al., 2003), (iii) oxidation of guanine leading to G to T transversions, (iv) deamination of adenine leading to A to G transitions (Moe et al., 2003), (v) oxidation of thymine leading to T to C transitions, (vi) incorporation of 8-oxo-G into the growing DNA strand opposite adenine followed by the replacement of A with C and the excision of 8-oxo-G leading to A to C transversions. Mono- and poly-functional enzymes involved in reparation of above-mentioned lesions are found in species from all three superkingdoms of life (Gros et al., 2002).

Second, translational selection, selection for optimal speed and accuracy of translation, should exert an influence on SCUB because preferred codons tend to correlate with the most common tRNAs (Ikemura, 1982), allowing for faster, yet accurate, codon recognition and translation of highly expressed genes (Curran and Yarus, 1989). Translation is the process by which ribosomes synthesize proteins in cells. The term “translational selection” refers to selection to optimize the translation process itself rather than selection acting on the functions of the proteins produced by translation (Ran and Higgs, 2012). One of the main pieces of evidence for translational selection is the observation that the choice of synonymous codons appears to be influenced by selection in many organisms. Synonymous changes in the gene do not affect the resulting protein but can affect the way that the mRNA is translated by the ribosome. The codon utilization scheme in all the organisms indicates the presence of translation selection as a major force in shaping codon usage (Pal et al., 2013). The speed of translation is one of the key factors on which translational selection can act. Speed has the direct benefit that the proteins required are produced faster, and the secondary benefit that if a given ribosome finishes translation of one sequence, it can begin work on another (Ran and Higgs, 2012). Hence, speeding up translation means that the same total protein production rate can be achieved with fewer ribosomes. The other important aspect of translational selection is accuracy. Occasional mis-pairings between codon and anticodon may occur during translation, leading to errors in the protein sequence, and even incorrect folding in the protein structure that could cause several neurodegenerative diseases (Drummond and Wilke, 2008). Thus, accurate translation of the sites that are evolutionarily conserved between species are particularly important for protein function should be particularly important, and the frequency of the most accurate codons should be higher at the conserved sites (Ran and Higgs, 2012). Our recent studies have indicated that SCUB could be crucial for understanding the etiology of central nervous system neurodegenerative diseases (CNSNDD) especially Alzheimer's disease (AD) as well as genetic factors. G and C ending codons were strongly biased in the coding sequences of the proteins related to AD as a result of genomic GC composition constraints, while codons that identified as translationally optimal in the major trend all end in C or G suggested that translational selection should also be taken into consideration additional to compositional constraints (Yang et al., 2010).

Although the causes of SCUB are still under discussion, factors that influence the codon usage patterns are: the GC content of the genome, especially the GC content at the third position of codons (Scapoli et al., 2009); concentrations of corresponding acceptor tRNA molecules (Higgs and Ran, 2008); the functions or hydrophilicity of proteins expressed by a gene (Zhang et al., 2009); gene expression levels; the amino acids composition of a protein (Lobry and Gautier, 1994); the structure of proteins (D'Onofrio et al., 2002); the mutational frequency and the method of mutation (Sueoka, 1992), etc. All these factors can be summarized as the influence of mutational pressure and translational selection (natural selection). Our results revealed that different archaeal virus species have different genomic G+C content which are ranging from 25% to 61%. This large variation of genomic G+C content of archaeal virus is different with that of other phages such as GC-rich Mycobacterium phage with genomic G+C content of 57–69% (Hassan et al., 2009) and AT-rich Staphylococcus phage (Bishal et al., 2012) with genomic G+C content of 30–37%. SIRV1 and SIRV2 in group 2 are the unenveloped, stiff-rod-shaped, linear dsDNA viruses isolated from Icelandic Sulfolobus, a novel virus family, the Rudiviridae (Prangishvili et al., 1999). Prangishvili and co-worker proposed that wild-type SIRV1 was unable to propagate in some hosts but surmounted this host range barrier by inducing a host response effecting extensive variation of the viral genome (Prangishvili et al., 1999). SIRV1 and SIRV2 preferred U- and A-ended codons whereas their host *Sulfolobus islandicus* with low AT % (Nayak, 2013), which could be due to cytosine deamination to uracil, similar to eukaryotic ssDNA viruses. Similarly, C to T transitions may be tolerated only at synonymous sites, resulting in an overabundance of thymine in the genomic sequences, and a corresponding preference for adenine in the coding sequence of SIRV1 and SIRV2. Here, mutational pressure may influence codon bias in SIRV1 and SIRV2 viruses. Meanwhile, Nayak has indicated the dominant role of mutational bias in codon usage pattern of *Sulfolobus islandicus'* genes based on the significant correlation between major trend of synonymous codon usage and GC3s.

Especially, ARV1 and AFV1 infecting Acidianus genus are different from other viruses in group 1, preferring A- and U-ended codons. The acidianus genus consists of acidothermophiles which grow optimally and slowly in the temperature range 65–95 °C and at pH 2–4 and belongs to the order sulfolobales (You et al., 2011; Kletzin, 1992; Vestergaard et al., 2005). Although the genome sequence and composition differ strongly from those of the Sulfolobus rudiviruses SIRV1 and SIRV2, they all carry the motif AATTTAGGAATTTAGGAATTT near the genome ends which may constitute a signal for the Holliday junction resolvase and DNA replication (Vestergaard et al., 2005). Similarly, C to T transitions may be tolerated only at synonymous sites, resulting in an overabundance of thymine in the genomic sequences, and a corresponding preference for adenine in the coding sequence of ARV1 and AFV1. Here, mutational pressure may influence codon bias in ARV1 and AFV1 viruses. Interestingly, bacteriophage KVP40, a double-stranded DNA phage belonging to a member of the *Myoviridae* family with an overall G+C content of 42.6%, was isolated from marine water and has been reported to infect eight *Vibrio* and one *Photobacterium* species. SCUB in KVP40 was determined to be AT-rich at the third codon positions, and their variations are dictated principally by both mutational bias and translational selection (Sau et al., 2007). Further analysis revealed that the RSCU of KVP40 is distinct from that of its Vibrio hosts, *Vibrio cholerae* and *V. parahaemolyticus*.

On the other hand, archaeal viruses are remarkably diverse in morphology and most of them represent unique morphotypes, some of which are still unclassified (Pina et al., 2011). Recently, metagenomics projects have revealed important information about the diversity and abundance of archaeal viruses in habitats populated by *Archaea* or led to the discovery of new archaeal

viruses PhiCh1 is a representative archaeal tailed dsDNA virus (Klein et al., 2012). It is a head-tail virus containing a contractible tail, thus belonging to the *Myoviridae* family. It infects the haloalkaliphilic Archaeon *Natrialba magadii* (Witte et al., 1997), which, to date, is its only known host. All of the previously described archaeal viruses have a double-stranded DNA (dsDNA) genome. *Halorubrum* pleomorphic virus 1 (HRPV-1), a newly characterized haloarchaeal virus, has a single-stranded DNA (ssDNA) genome (Pietilä et al., 2009). HRPV-1 and its host *Halorubrum* sp. were isolated from an Italian (Trapani, Sicily) solar saltern. Quantitative lipid comparison of HRPV-1 and its host revealed that HRPV-1 acquires lipids nonselectively from the host cell membrane, which is typical of pleomorphic enveloped viruses.

Particularly, PhiCh1 and HRPV1 infecting the kingdom Euryarchaeota are different from each other, except preferring C- and G-ended codons, involving the common 9 amino acids as follows: 7 amino acids prefer codons end with C, including including Asn (AAC), Asp (GAC), His (CAC), Tyr (UAC), Phe (UUC), Cys (UGC), Ile (AUC), and Leu (CUC), while 2 amino acids prefer codons end with G, such as Gln (CAG) and Glu (GAG). Besides, the rest amino acids in PhiCh1 still prefer G- or C-ended codons, such as Val (GUC), Gly (GGC), Thr (ACG), Ala (GCG), Lys (AAG), and Ser (UCC), as well as 2 amino acids preferring both G- and C-ended codons, including Pro (CCC/CCG) and Arg (CGC/CGG). However, except Arg, the rest amino acids mostly prefer codons ended with U in HRPV1, only Lys (AAA) prefer codons ended with A. It is well known that ssDNA is prone to rapid cytosine deamination to uracil (Frederico et al., 1990), which may explain the preference for U-ending codons in ssDNA HRPV1. Interestingly, there is a clear association between the number of synonymous codons encoding an amino acid (degeneracy level) and the preference for U or C at the third codon position in HRPV1, where U is preferred over C at the third codon position in amino acids encoded by four codons (i.e. Val, Gly, Pro, Thr, and Ala), while C is preferred over U in amino acids encoded by two and three codons (e.g. Phe, Tyr, His, Asp, Asn, and Ile). Wald et al. (2012) have found by a large-scale analysis to assess SCUB pattern of pyrimidine-ending codons in highly expressed genes in prokaryotes (using genes encoding ribosomal proteins of prokaryotes as proxy for highly expressed genes) that codon-pairs that encode two- and three-fold degenerate amino acids are biased towards C-ending codon while codons encoding four-fold degenerate amino acids are biased towards U-ending codon. This codon usage pattern is widespread in prokaryotes, and its strength is correlated with translational selection both within and between organisms, which could be emphasized in organisms where translational selection is operational in highly expressed genes. This bias is associated with an improved correspondence with the tRNA pool, avoidance of mis-incorporation errors during translation and moderate stability of codon-anticodon interaction, all consistent with more efficient translation. This observation that differences in tRNA abundance cannot explain SCUB is widespread in prokaryotes, where due to the striking absence of tRNAs containing A at the wobble position (Marck and Grosjean, 2002), synonymous codons ending with a pyrimidine residue (U or C) are translated by a single tRNA containing G at the wobble position (Wald et al., 2012). We took advantage of this phenomenon to explore the pattern of HRPV1's SCUB that seems independent of tRNA abundance, which is consistent with the conclusion of Wald et al. (2012), focusing on pyrimidine-ending codon. So, we propose that translational selection plays a leading role in HRPV1's SCUB and HRPV1 and this could be more close to prokaryotes, which is supported by the result of Marck and Grosjean, Archaea as an "intermediate domain" between Eukarya and Bacteria from the tRNomic point of view (Marck and Grosjean, 2002).

To study the codon usage variation among different archaeal virus genes, ENC and GC3S values of different archaeal virus genes were calculated. ENC values of different archaeal virus genes vary from 31.77 to 61, with a mean value of 46.68 and S.D. of 10.36. When comparing with other viruses or phages such as H5N1 influenza virus (mean ENC=50.91) (Ahn et al., 2006; Zhou et al., 2005), SARS-covs (mean ENC=44.45) (Gu et al., 2004), and foot-and-mouth virus (mean ENC=51.53) (Zhong et al., 2007), the ENC values for archaeal virus can be either high or low, indicating that not all archaeal virus have the same overall extent of SCUB. For instance, genes in group 1 have a average ENC value of 54.56 which means the seven archaeal virus species (such as SSVK1, SSV2, SSV1, SSVRH, SSV4, ARV1, and AFV1) have low level of SCUB, while genes in both group 2 and group 3 have average ENC value of 36.67 and 38.08, respectively, indicating a much higher extent of SCUB although genes in these two groups tend to use different codons as discussed above. As the selection-mutation-drift model (Paul and Wen-Hsiung, 1986; Bulmer, 1991) said, mutational pressure and translational selection are generally thought to be the main factors that account for SCUB in different organisms. Shackelton et al. (2006) have analyzed the sequenced vertebrate-infecting DNA virus and concluded that SCUB patterns are strongly related with genomic GC content. This work suggests that, in vertebrate-infecting DNA virus, genome-wide mutation, rather than natural selection, is the major factor that determines the SCUB. In this study, a similar conclusion is reached that mutational pressure is the main factor that affects SCUB in different archaeal virus groups due to the general correlation between base composition and SCUB. This concept is also verified by the highly negative correlation between axis1 values and GC3s and the result of ENC-plot (Fig. 2), demonstrating role of codon bias as an important determinant of codon usage.

Subsequently, natural selection, such as gene length, translational selection and gene function, can also account for SCUB in different organisms (Zhou et al., 2005). The phenomenon that genes with similar function although in different viral genomes are more likely to cluster together in COA is revealed in some published results (Gu et al., 2004; Das et al., 2006). In our investigation, it is clear that gene function, albeit with smaller effects, also influence SCUB among these viral genes. Several researches indicated that hydropathy of each mimivirus' gene, aromaticity and cysteine content are mostly account for the variation of amino acid usage in mimivirus and foot-and-month disease virus (Zhong et al., 2007; Sau et al., 2006). In our study, the significant positive correlation between aromaticity of each hypothetical polyprotein and SCUB was found ($r=0.733$, $P<0.01$). Moreover, the cluster tree generated by the hierarchical clustering method based on the variation in RSCU values among archaeal virus' genes showed that the 26 genes were divided into 3 distinct sublineages, corresponding to the three groups (Fig. 4). So, genomic GC content is a dominant factor that affected the classification of archaeal viruses and gene function may also play important role in determining the classification at the level of sublineage although other factors may also involve in.

Furthermore, previous report suggests that CpG under-representation can affect codon usage preference in RNA and small DNA viruses (Karlin et al., 1990). Our results show that codon usage in archaeal virus can also be highly influenced by biases in dinucleotide frequencies. As an example, all CpG containing codons are largely suppressed in genes of group 2. As to group 3, not all codons containing UpC and GpA are over-used. This may be due to the fact that the only gene in HRPV1 with 54% GC content are also included in group 3, which may play an influence on the overall frequencies of dinucleotide. These results can also be explained by mutational pressure. It is well known that the global methylation pattern is a key feature of the methylation landscape of the human genome. Most of the gene bodies and intergenic sequences are globally methylated with the exception of regions called CpG islands (CGIs). CGIs are often unmethylated, but there are increasing number of CGIs being reported to be methylated in normal tissues (Fan and Zhang, 2009). In mammals, the methylated form of cytosine (5-methylcytosine) is hypermutable. 5-methylcytosine is formed by the enzyme DNA methyltransferase operating

on a cytosine occurring immediately 5′ of a guanine. One effect of methylation is to increase the rate of spontaneous deamination of 5-methylcytosine to form thymine. It has been estimated that transitions in the methylated CpG dinucleotide occur 8–16 times faster than non-CpG transitions (Gaffney and Keightley, 2008). The deficit of CpG dinucleotides in the gene coding sequences of archaeal virus (especially SIRV1 and SIRV2) is largely attributed to the hypermutability of methylated CpGs to UpGs (or CpAs in the complementary strand). Since, CpGs in CGIs are often unmethylated and their mutations are rare, the deficiency of CpG can be considered an influence of mutation. Our results also showed that the synonymous SCUB in archaeal virus' genes of group 3 are ordered by the third codon position as follows: C > G > U > A, which is in accord with Zeeberg's results (Zeeberg, 2012). This appears to be a manifestation of an evolutionary strategy for placement of genes in regions of the genome with a GC content that relates synonymous SCUB and protein folding. Remarkably, collections of Clusters of Orthologous Genes (COGs) for Archaea have described evolutionary reconstructions to reveal general trends in the evolution of Archaea and performed maximum likelihood reconstruction of the genome content of archaeal ancestral forms and gene gain and loss events in archaeal evolution (Wolf et al., 2012). This reconstruction showed that the last common ancestor of the extant Archaea was an organism of greater complexity than most of the extant archaea, probably with over 2500 protein-coding genes. The subsequent evolution of almost all archaeal lineages was apparently dominated by gene loss resulting in genome streamlining. Overall, gene losses are estimated to outnumber gene gains at least 4 to 1 in the evolution of Archaea. Analysis of specific patterns of gene gain in Archaea shows that gene exchange between major groups of Archaea appears to be largely random while the conserved core of archaeal genes appears to be stabilizing. Similarly, SCUB in AD and other neurodegenerative diseases indicated that GC-rich codons are mainly in charge of forming contracted conformation, especially the first nucleotide of codons plays a dominant role in translating the genomic GC signature into protein sequences and structures (Yang et al., 2010). Our previous research has revealed that conformation biases of amino acids are present in natural proteins and the corresponding biases of codons show an evident tendency in protein folding (Yang et al., 2006).

Eventually, viruses with well characterized hosts should experience translational selection to match the SCUB of their hosts, as this should allow for faster translation of highly expressed viral genes, and consequently more rapid viral replication (Cardinale et al., 2013). However, surveys examining viral SCUB have indicated that not all viruses are equally able to match their hosts' codon preferences, and that this may be correlated with viral genomic architecture (Jenkins and Holmes, 2003). For instance, dsDNA coliphages were significantly better matched to *Escherichia coli*'s SCUB than ssDNA coliphages, because ssDNA phages had a preference for NNT codons, regardless of the hosts' preferred codon usage (Cardinale and Duffy, 2011). Plant virus only infect monocots host but also infect eudicot hosts, such as the arthropod-vectored plant viruses, namely the positive sense ssRNA genus *Potyvirus* and family *Luteoviridae,* and the ssDNA family *Geminiviridae* (Cardinale et al., 2013). Monocots usually tend to have GC biased genes (53–56%), while eudicot genes generally have lower GC content (40–45%) (Wang and Roossinck, 2006). Monocots exclusively prefer G- and C-ending codons, while eudicots prefer a combination of G- and T-ending codons in their most highly expressed genes (Cardinale et al., 2013). These divergent hosts allow the strength of translational selection pressures to be compared among related viruses. All potyviruses had somewhat similar codon preferences, independent of host: monocot- and eudicot-infecting potyviruses both generally preferred A- and T-ending codons, while Luteoviruses, both monocot- and eudicot-infecting, tended to favor NNC codons (Cardinale et al., 2013). ssDNA dicot-infecting begomovirus CP genes exhibited a strong preference for NNT codons, while its Rep sequences strongly favored NNA codons. Begomovirus genomes are ambisense; genes are encoded in the coding and complimentary sense (Gutierrez, 1999). The coding sequence of the Rep gene is complimented on the virion strand. As a consequence, third positions in Rep gene are present as the first base of anti-codons in the single-stranded viral genome. Therefore, these findings indicate begomovirus genomes are enriched for thymine at synonymous sites in both the CP ORF (with T-ending codons) and Rep ORF (with T-beginning anticodons) (Cardinale et al., 2013). So, the codon preferences of ssRNA (luteoviruses and potyviruses) and ssDNA (geminiviruses) plant viruses could not directly relate with genomic base composition and translational selection, but constraints such as genomic architecture and secondary structure can and do influence codon usage in plant viruses. Additionally, our results indicate SIRV1 and SIRV2 preferred U- and A-ended codons which is not similar to their host *Sulfolobus islandicus* with low AT % (Nayak, 2013), which could be due to cytosine deamination to uracil, similar to eukaryotic ssDNA viruses. Similarly, ARV1 and AFV1 infecting Acidianus genus preferred A- and U-ended codons, which could be because of C to T transitions only at synonymous sites resulting in an overabundance of thymine in the genomic sequences and a corresponding preference for adenine in the coding sequence of ARV1 and AFV1. Here, mutational pressure may influence codon bias in SIRV1, SIRV2, ARV1 and AFV1 viruses. Particularly, PhiCh1 and HRPV1 infecting the kingdom Euryarchaeota are different from each other, except common 9 amino acids as follows: 6 two- and three-fold degenerate amino acids preferring C-ended codons, such as Phe, Tyr, His, Asp, Asn, and Ile, as well as six-fold amino acid Leu (CUC), and 2 two-fold amino acids preferring G-ended codons, such as Gln and Glu. 5 four-fold degenerate amino acids (i.e. Val, Gly, Pro, Thr, and Ala) prefer U-ended codons in HRPV1, but prefer G- or C-ended codons in PhiCh1. The conclusion that ssDNA is prone to rapid cytosine deamination to uracil (Frederico et al., 1990) could explain the preference for U-ending codons in ssDNA HRPV1, which supported the result of Wald et al. based on a large-scale analysis to assess SCUB pattern of pyrimidine-ending codons in highly expressed genes in prokaryotes (Wald et al., 2012). This is owing to the fact that the striking absence of tRNAs containing A at the wobble position (Marck and Grosjean, 2002) result in translation of synonymous codons ending with a pyrimidine residue (U or C) by a single tRNA containing G at the wobble position (Wald et al., 2012). We report that translational selection plays a leading role in HRPV1's SCUB and HRPV1 can be more close to the prokaryotes in some properties.

0In conclusion, our study reveals the heterogeneity of synonymous codon usage among different archaeal virus species with different genomic GC contents, and mutational pressure is the main factor that influences SCUB. Other factors, such as dinucleotide composition, aromaticity and gene function also affects codon usage variation although they have fewer influences than mutational pressure. Mutational pressure may influence SCUB in SIRV1, SIRV2, ARV1, AFV1, and PhiCh1 viruses, whereas translational selection could play a leading role in HRPV1's SCUB. The results we reported not only can offer an insight into the codon usage pattern of archaeal virus and subsequently the possible relationship between archaeal virus and its host, but also is useful in understanding the evolution of archaeal virus and its gene classification, and more helpful to explore the origin of life and the evolution of biology.

Acknowledgments

This work is supported by Training project of scientific talent of National Natural Science Foundation of China (J1103512 and J1210026), Specialized Research Fund for the doctoral program of higher education (for PhD Advisors), the Ministry of Education of the People's Republic of China (20120091110038), and National Key Technology R&D Program (2008BAI51B01).

Appendix A. Supplementary materials

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2014.03.022.

References

Ackermann, H.W., 2007. 5500 phages examined in the electron microscope. Arch. Virol. 152, 227–243.

Ahn, I., Jeong, B.J., Bae, S.E., Jung, J., Son, H.S., 2006. Genomic analysis of influenza A viruses, including avian flu (H5N1) strains. Eur. J. Epidemiol. 21, 511–519.

Albrecht-Buehler, G., 2006. Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. Proc. Natl. Acad. Sci. USA 103, 17828–17833.

Angellotti, M.C., Bhuiyan, S.B., Chen, G., Wan, X.F., 2007. CodonO: codon usage bias analysis within and across genomes. Nucleic Acids Res. 35, W132–W136.

Bettstetter, M., Peng, X., Garrett, R.A., Prangishvili, D., 2003. AFV1, a novel virus infecting hyperthermophilic archaea of the genus acidianus. Virology 315, 68–79.

Bishal, A.K., Saha, S., Sau, K., 2012. Synonymous codon usage in forty staphylococcal phages identifies the factors controlling codon usage variation and the phages suitable for phage therapy. Bioinformation 8, 1187–1194.

Bulmer, M., 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129, 897–907.

Cardinale, D.J., DeRosa, K., Duffy, S., 2013. Base composition and translational selection are insufficient to explain codon usage bias in plant viruses. Viruses 5, 162–181.

Cardinale, D.J., Duffy, S., 2011. Single-stranded genomic architecture constrains optimal codon usage. Bacteriophage 1, 219–224.

Curran, J.F., Yarus, M., 1989. Rates of aminoacyl-trna selection at 29 sense codons in vivo. J. Mol. Biol. 209, 65–77.

Das, S., Paul, S., Dutta, C., 2006. Synonymous codon usage in adenoviruses: influence of mutation, selection and protein hydropathy. Virus Res. 117, 227–236.

D'Onofrio, G., Ghosh, T.C., Bernardi, G., 2002. The base composition of the genes is correlated with the secondary structures of the encoded proteins. Gene 300, 179–187.

Drummond, D.A., Wilke, C.O., 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134, 341–352.

Drummond, D.A., Wilke, C.O., 2009. The evolutionary consequences of erroneous protein synthesis. Nat. Rev. Genet. 10, 715–724.

Fan, S., Zhang, X., 2009. CpG island methylation pattern in different human tissues and its correlation with gene expression. Biochem. Biophy. Res. Commun. 383, 421–425.

Frederico, L.A., Kunkel, T.A., Shaw, B.R., 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry 29, 2532–2537.

Gaffney, D.J., Keightley, P.D., 2008. Effect of the assignment of ancestral CpG state on the estimation of nucleotide substitution rates in mammals. BMC Evol. Biol. 8, 265.

Grantham, R., Gautier, C., Gouy, M., 1980. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res. 8, 1893–1912.

Grocock, R.J., Sharp, P.M., 2002. Synonymous codon usage in *Pseudomonas aeruginosa* PA01. Gene 289, 131–139.

Gros, L., Saparbaev, M.K., Laval, J., 2002. Enzymology of the repair of free radicals-induced DNA damage. Oncogene 21, 8905–8925.

Gu, W., Zhou, T., Ma, J., Sun, X., Lu, Z., 2004. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. Virus Res. 101, 155–161.

Gutierrez, C., 1999. Geminivirus DNA replication. Cell. Mol. Life Sci. 56, 313–329.

Haring, M., Peng, X., Brqgger, K., Rachel, R., Stetter, K.O., Garrett, R.A., Prangishvili, D., 2004. Morphology and genome organisation of the virus PSV of the hyperthermophilic archaeal genera pyrobaculum and thermoproteus: a novel virus family, the Globuloviridae. Virology 323, 233–242.

Hassan, S., Mahalingam, V., Kumar, V., 2009. Synonymous codon usage analysis of thirty two mycobacteriophage genomes. Adv Bioinform., 316936

Higgs, P.G., Ran, W., 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol. Biol. Evol. 25, 2279–2291.

Holm, L., 1986. Codon usage and gene expression. Nucleic Acids Res. 14, 3075–3087.

Ibba, M., Soll, D., 1999. Quality control mechanisms during translation. Science 286, 1893–1897.

Ikemura, T., 1982. Correlation between the abundance of yeast transfer rnas and the occurrence of the respective codons in protein genes: differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer rnas. J. Mol. Biol. 158, 573–597.

Jenkins, G.M., Holmes, E.C., 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. 92, 1–7.

Kahali, B., Basak, S., Ghosh, T.C., 2007. Reinvertigating the codon and amino acid usage of *S. Cerevisiae* genome: a new insight from protein secondary structure analysis. Biochem. Biophys. Res. Commun. 354, 693–699.

Karlin, C., Burge, C., 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11, 283–290.

Karlin, S., Blaisdell, B.E., Schachtel, G.A., 1990. Contrasts in codon usage of latent versus productive genes of Epstein–Barr virus: data and hypotheses. J. Virol. 64, 4264–4273.

Karlin, S., Ladunga, I., Blaisdell, B.E., 1994. Heterogeneity of genomes: measures and values. Proc. Natl. Acad. Sci. USA 91, 12837–12841.

Karlin, S., Burge, C., 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11, 283–290.

Karline, C., Marazek, J., 1996. What drives codon choices in human genes? Mol. Biol. 262, 459–472.

Khrustalev, V.V., Barkovsky, E.V., 2010. Study of completed archaeal genomes and proteomes: hypothesis of strong mutational AT pressure existed in their common predecessor. Genomics Proteomics Bioinform. 8, 22–32.

Klein, R., Baranyi, U., Rossler, N., Greineder, B., Scholz, H., Witte, A., 2002. *Natrialba magadii* virus fCh1: first complete nucleotide sequence and functional organization of a virus infecting a haloalkaliphilic archaeon. Mol. Microbiol. 45, 851–863.

Klein, R., Rossler, N., Iro, M., Scholz, H., Witte, A., 2012. Haloarchaeal myovirus φCh1 harbours a phase variation system for the production of protein variants with distinct cell surface adhesion specificities. Mol. Microbiol. 83, 137–150.

Kletzin, A., 1992. Molecular characterization of the sor gene, which encodes the sulfur oxygenase/reductase of the thermoacidophilic archaeum desulfurolobus ambivalens. J. Bacteriol. 174, 5854–5859.

Kong, S.G., Fan, W.L., Chen, H.D., Hsu, Z.T., Zhou, N., Zheng, B., Lee, H.C., 2009. Inverse symmetry in complete genomes and whole-genome inverse duplication. PLoS One 4, e7553.

Kurland, C.G., 1991. Codon bias and gene expression. FEBS Lett. 285, 165–169.

Lavner, Y., Kotlar, D., 2005. Codon bias as a factor in regulating expression via translation rate in the human genome. Gene 345, 127–138.

Liu, Y.X., Li, S., Yang, J., 2010. Analysis of synonymous codon usages in the gene coding sequences of diabetes related proteins. J. Nanjing Univ. 46, 114–119.

Lobry, J.R., Gautier, C., 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res. 22, 3174–3180.

Ma, J., Campbell, A., Karlin, S., 2002a. Correlations between shine-dalgarno sequences and gene features such as predicted expression levels and operon structures. J. Bacteriol. 184, 5733–5745.

Ma, J., Zhou, T., Gu, W., Sun., X., Lu, Z., 2002b. Cluster analysis of the codon use frequency of MHC genes from different species. Biosystems 65, 199–207.

Marck, C., Grosjean, H., 2002. tRNomics: analysis of tRNA genes from 50 genomes of eukarya, archaea, and bacteria reveals anticodon-sparing strategies and domain-specific features. RNA 8, 1189–1232.

Moe, A., Ringvoll, J., Nordstrand, L.M., Eide, L., Bjørås, M., Seeberg, E., Rognes, T., Klungland, A., 2003. Incision at hypoxanthine residues in DNA by a mammalian homologue of the *Escherichia coli* antimutator enzyme endonuclease V. Nucleic Acids Res. 31, 3893–3900.

Mooers, A.O., Holmes, E.C., 2000. The evolution of base composition and phylogenetic inference. Trends Ecol. Evol. 15, 365–369.

Mrázek, J., Karlin, S., 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. USA 95, 3720–3725.

Nakamura, T., Suyama, A., Wada, A., 1991. Two types of linkage between codon usage and gene-expression levels. FEBS Lett. 289, 123–125.

Nayak, K.C., 2013. Comparative genome sequence analysis of sulfolobus acidocaldarius and 9 other isolates of its genus for factors influencing codon and amino acid usage. Gene 513, 163–173.

Necşulea, A., Lobry, J.R., 2007. A new method for assessing the effect of replication on DNA base composition asymmetry. Mol. Biol. Evol. 24, 2169–2179.

Niu, D.K., Lin, K., Zhang, D.Y., 2003. Strand compositional asymmetries of nuclear DNA in eukaryotes. J. Mol. Evol. 57, 325–334.

Pal, A., Mukhopadhyay, S., Bothra, A.K., 2013. Statistical analysis of pentose phosphate pathway genes from eubacteria and eukarya reveals translational selection as a major force in shaping codon usage pattern. Bioinformation 9, 349–356.

Palidwor, G.A., Perkins, T.J., Xia, X., 2010. A general model of codon bias due to GC mutational bias. PLoS One 5, e13431.

Palm, P., Schleper, C., Grampp, B., Yeats, S., McWilliam, P., Reiter, W.D., Zillig, W., 1991. Complete nucleotide sequence of the virus SSV1 of the archaebacterium sulfolobus shibatae. Virology 185, 242–250.

Pan, T., Li, D., Luo., M., Tang, F., 2009. Analysis of synonymous codon usage in classical swine fever virus. Virus Genes 38, 104–112.

Paul, M.S., Wen-Hsiung, L., 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28–38.

Peng, X., 2008. Evidence for the horizontal transfer of an integrase gene from a fusellovirus to a pRN-like plasmid within a single strain of sulfolobus and the implications for plasmid survival. Microbiology 154, 383–391.

Peng, X., Blum, H., She, Q., Mallok, S., Brügger, K., Garrett, R.A., Zillig, W., Prangishvili, D., 2001. Sequences and replication of genomes of the archaeal rudiviruses SIRV1 and SIRV2: relationships to the archaeal lipothrixvirus SIFV and some eukaryal viruses. Virology 291, 226–234.

Pietilä, M.K., Roine, E., Paulin, L., Kalkkinen, N., Bamford, D.H., 2009. An ssDNA virus infecting archaea: a new lineage of viruses with a membrane envelope. Mol. Microbiol. 72, 307–319.

Pina, M., Bize, A., Forterre, P., Prangishvili, D., 2011. The archeoviruses. FEMS Microbiol. Rev. 35, 1035–1054.

Prangishvili, D., Arnold, H.P., Gotz, D., Ziese, U., Holz, I., Kristjansson, J.K., Zillig, W., 1999. A novel virus family, the rudiviridae: structure, virus-host interactions

and genome variability of the sulfolobus viruses SIRV1 and SIRV2. Genetics 152, 1387–1396.
Prangishvili, D., Forterre, P., Garrett, R.A., 2006. Viruses of the archaea: a unifying view. Nat. Rev. Microbiol. 4, 837–848.
Prangishvili, D., Garrett, R.A., 2004. Exceptionally diverse morphotypes and genomes of crenarchaeal hyperthermophilic viruses. Biochem. Soc. Trans. 32, 204–208.
Precup, J., Parker, J., 1987. Missense misreading of asparagine codons as a function of codon identity and context. J. Biol. Chem. 262, 11351–11355.
Ran, W., Higgs, P.G., 2012. Contributions of speed and accuracy to translational selection in bacteria. PLoS One 7, e51652.
Satapathy, S.S., Dutta, M., Buragohain, A.K., Ray, S.K., 2012. Transfer RNA gene numbers may not be completely responsible for the codon usage bias in asparagine, isoleucine, phenylalanine, and tyrosine in the high expression genes in bacteria. J. Mol. Evol. 75, 34–42.
Sanchez, J., 2011. 3-base periodicity in coding DNA is affected by intercodon dinucleotides. Bioinformation 6, 327–329.
Sanchez, J., Jose, M.V., 2002. Analysis of bilateral inverse symmetry in whole bacterial chromosomes. Biochem. Biophys. Res. Commun. 299, 126–134.
Sau, K., Gupta, S.K., Sau, S., Mandal, S.C., Ghosh, T.C., 2006. Factors influencing sysnonymous codon and amino acid usage biases in mimivirus. Biosystems 85, 107–113.
Sau, K., Gupta, S.K., Sau, S., Ghosh, T.C., 2005. Synonymous codon usage bias in 16 Staphylococcus aureus phages: implication in phage therapy. Virus Res. 113, 123–131.
Sau, K., Gupta, S.K., Sau, S., Mandal, S.C., Ghosh, T.C., 2007. Studies on synonymous codon and amino acid usage biases in the broad-host range bacteriophage KVP40. J Microbiol. 45, 58–63.
Scapoli, C., Bartolomei, E., De Lorenzi, S., Carrieri, A., Salvatorell, G., Rodriguez-Larralde, A., Barrai, I., 2009. Codon and aminoacid usage patterns in mycobacteria. J. Mol. Microbiol. Biotechnol. 17, 53–60.
Shackelton, L.A., Parrish, C.R., Holmes, E.C., 2006. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. J. Mol. Evol. 62, 551–563.
Snyder, J.C., Samson, R.Y., Brumfield, S.K., Bell, S.D., Young, M.J., 2013. Functional interplay between a virus and the ESCRT machinery in archaea. Proc. Natl. Acad. Sci. USA 110, 10783–10787.
Stedman, K.M., She, Q., Phan, H., Arnold, H.P., Holz, I., Garrett, R.A., Zillig, W., 2003. Relationship between fuselloviruses infecting the extremely thermophilic archaeon sulfolobus: SSV1 and SSV2. Res. Microbiol. 154, 295–302.
Sueoka, N., 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA 48, 582–592.
Sueoka, N., 1992. Directional mutation pressure, selective constraints, and genetic equilibria. J. Mol. Evol. 34, 95–114.
Sueoka, N., 1998. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA 85, 2653–2657.
Sueoka, N., 2002. Wide intra-genomic G+C heterogeneity in human and chicken is mainly due to strand-symmetric directional mutation pressures: dGTP-oxidation and symmetric cytosine-deamination hypotheses. Gene 300, 141–154.
Tolstrup, N., Sensen, C.W., Garrett, R.A., Clausen, I.G., 2000. Two different and highly organized mechanisms of translation initiation in the archaeon sulfolobus solfataricus. Extremophiles 4, 175–179.
Varenne, S., Buc, J., Lloubes, R., Lazdunski, C., 1984. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J. Mol. Biol. 180, 549–576.
Vestergaard, G., Häring, M., Peng, X., Rachel, R., Garrett, R.A., Prangishvili, D., 2005. A novel rudivirus, ARV1, of the hyperthermophilic archaeal genus acidianus. Virology 336, 83–92.
Wald, N., Alroy, M., Botzman, M., Margalit, H., 2012. Codon usage bias in prokaryotic pyrimidine-ending codons is associated with the degeneracy of the encoded amino acids. Nucleic Acids Res. 40, 7074–7083.
Wang, L., Roossinck, M.J., 2006. Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants. Plant Mol. Biol. 61, 699–710.
Wiedenheft, B., Stedman, K., Roberto, F., Willits, D., Gleske, A.K., Zoeller, L., Snyder, J., Douglas, T., Young, M., 2004. Comparative genomic analysis of hyperthermophilic archaeal *fuselloviridae* viruses. J. Virol. 78, 1954–1961.
Witte, A., Baranyi, U., Klein, R., Sulzner, M., Luo, C., Wanner, G., Krüger, D.H., Lubitz, W., 1997. Characterization of *Natronobacterium magadii* phage fCh1, a unique archaeal phage containing DNA and RNA. Mol. Microbiol. 23, 603–616.
Woese, C.R., Kandler, O., Wheelis, M.L., 1990. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. Proc. Natl. Acad. Sci. USA 87, 4576–4579.
Wolf, Y.I., Makarova, K.S., Yutin, N., Koonin, E.V., 2012. Updated clusters of orthologous genes for archaea: a complex ancestor of the archaea and the byways of horizontal gene transfer. Biol. Direct 7, 46.
Wright, F., 1990. The 'effective number of codons' used in a gene. Gene 87, 23–29.
Yang, J., Dong, X.C., Leng, Y., 2006. Conformation biases of amino acids based on tripeptide microenvironment from PDB database. J. Theor. Biol. 240, 374–384.
Yang, J., Zhu, T.Y., Jiang, Z.X., Chen, C., Wang, Y.L., Zhang, S., Jiang, X.F., Wang, T.T., Wang, L., Xia, W.H., Li, L., Chen, J.J., Wang, J.Y., Wang, W.W., Zheng, W.J., 2010. Codon usage biases in Alzheimer's disease and other neurodegenerative diseases. Protein Pept. Lett. 17, 630–645.
Yoon, J.H., Iwai, S., O'Connor, T.R., Pfeifer, G.P., 2003. Human thymine DNA glycosylase (TDG) and methyl-CpG-binding protein 4 (MBD4) excise thymine glycol (Tg) from a Tg:G mispair. Nucleic Acids Res. 31, 5399–5404.
You, X.Y., Liu, C., Wang, S.Y., Jiang, C.Y., Shah, S.A., Prangishvili, D., She, Q., Liu, S.J., Garrett, R.A., 2011. Genomic analysis of acidianus hospitalis W1 a host for studying crenarchaeal virus and plasmid life cycles. Extremophiles 15, 487–497.
Zaher, H.S., Green, R., 2009. Quality control by the ribosome following peptide bond formation. Nature 457, 161–166.
Zeeberg, B., 2012. Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes. Genome Res. 12, 944–955.
Zhang, J., Wang, M., Liu, W.Q., Zhou, J.H., Chen, H.T., Ma, L.N., Ding, Y.Z., Gu, Y.X., Liu, Y.S., 2011. Analysis of codon usage and nucleotide composition bias in polioviruses. Virol. J. 8, 146.
Zhang, Q., Zhao, S., Chen, H., Liu, X., Zhang, L., Li, F., 2009. Analysis of the codon use frequency of AMPK family genes from different species. Mol. Biol. Rep. 36, 513–519.
Zhao, K.N., Liu, W.J., Frazer, I.H., 2003. Codon usage bias and A+T content variation in human papillomavirus genomes. Virus Res. 98, 95–104.
Zhong, J., Li, Y., Zhao, S., Liu, S., Zhang, Z., 2007. Mutation pressure shapes codon usage in the GC-rich genome of foot-and-mouth disease virus. Virus Genes 35, 767–776.
Zhou, J., Liu, W.J., Peng, S.W., Sun, X.Y., Frazer, I., 1999. Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. J. Virol. 73, 4972–4982.
Zhou, T., Gu, W., Ma, J., Sun, X., Lu, Z., 2005. Analysis of synonymous codon usage in H5N1 virus and other *influenza* A viruses. Biosystems 81, 77–86.