Biochemical and Biophysical Research Communications 430 (2013) 1344–1348

Contents lists available at SciVerse ScienceDirect

Biochemical and Biophysical Research Communications

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

Mini Review

Analysis of codon usage patterns in *Taenia pisiformis* through annotated transcriptome data

Lin Chen [a], Tianfei Liu [b], Deying Yang [a], Xiang Nong [a], Yue Xie [a], Yan Fu [a], Xuhang Wu [a], Xing Huang [a], Xiaobin Gu [a], Shuxian Wang [a], Xuerong Peng [c], Guangyou Yang [a,*]

[a] *Department of Parasitology, College of Veterinary Medicine, Sichuan Agricultural University, Ya'an 625014, China*
[b] *Institute of Animal Science, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China*
[c] *College of Life and basic Science, Sichuan Agricultural University, Ya'an 625014, China*

ARTICLE INFO

Article history:
Received 28 November 2012
Available online 23 December 2012

Keywords:
Taenia pisiformis
Transcriptome
Codon usage bias
Correspondence analysis
Optimal codon

ABSTRACT

Taenia pisiformis (Cestoidea; Cyclophyllidea; Taeniidae) tapeworms infect the small intestine of canids and felines, such as dogs and foxes. Synonymous codon usage in *T. pisiformis* was examined through 8118 reconstructed annotations of transcriptome sequences. The mean value of GC content for the reconstructed genes was 49.48%. Twenty-four codons were determined as "optimal codons". Approximately all translational optimal codons (except CGU) ended on G or C. The gene positions on the primary axis were strongly positively correlated with GC content at the third codon positions and GC content of individual genes. At the same time, the gene expression level assessed by the CAI, the hydrophobicity and aromaticity of encoded proteins were correlated with the GC content at the third codon positions and the effective number of codons (ENC), respectively. We infer that the gene expression level, the hydrophobicity and the aromaticity of the encoded proteins also influenced codon usage in *T. pisiformis*. Knowledge of the codon usage pattern in *T. pisiformis* can improve our understanding of the mechanisms of biased usage of synonymous codons and can help in selecting appropriate host expression systems for potential vaccine genes of *T. pisiformis*.

1. Introduction

Due to the degeneracy of the genetic code, most amino acids (except methionine and tryptophane) are encoded by more than one codon. Synonymous codons are often used with very different frequencies both within and between genomes [1,2]. Many factors have been reported to influence codon usage in various organisms. Compositional constraints and translational selection are thought to be the main factors that account for codon usage variation among genes in different organisms [3].

Studies of synonymous codon usage can improve our understanding of the mechanisms of synonymous codon-biased usage [4], the selection of appropriate host expression systems [4,5], the design of degenerate primers [6], gene prediction from genomic sequences [7], and protein functional classification [8]. In addition, profiles of synonymous codon usage can reveal information about the molecular evolution of individual genes and provide data to train genome-specific gene recognition algorithms, which detect protein-coding regions in uncharacterized genomic DNA [9].

The adult stages of *Taenia pisiformis* (Cestoidea; Cyclophyllidea; Taeniidae) parasitize in the small intestine of canids and felines. *T. pisiformis* can cause significant health problems and even death to their hosts [10]. At present, synonymous codon usage biases have been determined in numerous organisms. Only 27 sequences of *T. pisiformis* have been registered at NCBI as of April 9th, 2012, and only four of these are complete nuclear gene CDSs: TP14 8 kDa glycoprotein (GenBank: GU321333.1), cathepsin L-like cysteine protease (GenBank: JF798507.1), cysteine protease (GenBank: JF718743.1), and 18S ribosomal RNA gene (JQ609339.1). Therefore, the available molecular biological information regarding *T. pisiformis* on NCBI is limited. There have also been few studies of codon and nucleotide usage biases in *T. pisiformis*. In this study, we investigated the codon usage profile of *T. pisiformis* from annotations of the transcriptome using a multivariate statistical analysis. Knowledge of the codon usage pattern in *T. pisiformis* can provide a basis for understanding the mechanisms of biased usage of synonymous codons and for selecting appropriate host expression systems to improve the expression of vaccine genes of *T. pisiformis*.

2. Materials and methods

2.1. Transcriptome data

A total of 72,957 non-redundant consensus sequences from adult *T. pisiformis* were retrieved using a high-throughput

* Corresponding author. Fax: +86 835 2885302.
E-mail address: guangyou1963@yahoo.com.cn (G. Yang).

0006-291X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2012.12.078

sequencing (RNA-seq) method [10]. Unigenes were aligned with clean reads of *T. pisiformis*. The results showed that 99.79% (7802/72,957) unigenes mapped to clean reads. This suggested that the quality of the adult *T. pisiformis* unigenes data was reliable.

2.2. Prediction of coding sequences in transcriptome data

Based on a sequence similarity search with known proteins, a total of 68,945 unigenes were identified. Annotations of 68,945 unigenes using the NCBI non-redundant (Nr) protein database, the Swiss-prot protein database, the Clusters of Orthologous Groups of proteins (COGs) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) protein databases (with an *E*-value cut-off of 1×10^{-5}) showed that 25,701, 19,564, 7760 and 15,920 unigene consensus sequences had a high similarity with the known genes of existing species, respectively. Among the annotated unigenes, 25,633 coding sequences (CDS) were obtained by the BLASTx algorithm, where one unigene corresponded to one CDS. All CDSs were analyzed using the FrameDP software, which has the ability to self-train directly on EST clusters instead of requiring curated cDNA sets to train the underlying ESTScan and DECODER software.

According to the annotation information and published literature, all mitochondrial genes (29 CDSs) and ribosomal genes (59 CDSs) were confirmed in the data. Putative ribosomal genes were confirmed manually. Compared with the Database of Essential Genes (DEG) [11], 39 CDSs were confirmed manually to be essential genes. We excluded these 29 mitochondrial genes and genes with gaps (786 CDSs) from the analysis, as we only considered the codon usage of nuclear protein CDSs.

To improve the quality of sequences and minimize sampling errors, only CDS sequences longer than or equal to 450 bp were used for this study. A total of 8118 reconstructed genes met the above criteria and were selected for further analysis. It should be noted that some of the sequences were partial even though they are referred to as “genes”.

2.3. Indices of codon usage bias

Different codon indices were calculated: relative synonymous codon usage (RSCU), the effective number of codons (ENC), the codon adaptation index (CAI), and the frequency of GC content at the third codon positions (GC3s). RSCU values were used to study the overall synonymous codon usage variation among the genes. A GC3s value is the frequency of GC content at the third synonymously variable coding position (excluding methionine, tryptophane and the three stop codons). It is a good indicator of the extent of base composition bias, and the expected ENC values from GC3s were computed.

2.4. Determination of optimal codons

Based on the calculated CAI values, 5% of the total genes with extremely high and low CAI values were regarded as the high and low datasets, respectively. Codon usage was compared using the Chi-squared contingency test of the two groups, and codons whose frequency of usage were significantly higher ($P < 0.01$) in highly expressed genes than in genes with low level of expression were defined as the optimal codons.

2.5. Correspondence analysis of codon usage

The relationships between variables and samples can be explored using a multivariate statistical analysis. Correspondence analysis (COA) is a widely used approach in codon usage analysis, and was used to study the variation trends among CDSs. CodonW 1.4.2 software was used to analyze the indices of codon usage. Statistical analysis was performed with SPSS12.0, Origin 8.0 and Excel 2003.

3. Results

3.1. Nucleotide contents of genes

The GC contents of the reconstructed genes varied from 27.0% to 69.8% with a standard deviation (SD) of 3.90. The GC contents of the reconstructed genes were mainly distributed between 40% and 60% (Fig. 1). To understand the nucleotide distribution, we investigated the GC and GC3s content, and the effective number of codon ENC of the reconstructed genes (Table 1). The mean value of GC contents for the reconstructed genes was 49.84%, indicating that the nucleotides A and T were almost equally distributed to G and C. The ENC values varied from 27.55 to 61, with a mean value of 55.78 and a SD of 4.55.

4. Preferential codon usage

The overall RSCU values of 59 sense codons in *T. pisiformis* are shown in Table 2. Thirty-two codons, including GCU, GCC, and CGU, were frequently used codons, which were significantly more frequent among the highly expressed genes. Approximately half (17/32) of the frequently used codons are ended on G or C. These results suggest that compositional limitations often play an integral role in the codon usage pattern of *T. pisiformis*.

4.1. Codon usage analysis

Plotting ENC and GC3s is an effective way to explore the main features of codon usage among genes [12]. The ENC value of each reconstructed gene was plotted against its corresponding GC3 (Fig. 2). The solid line represents the expected positions of genes whose codon usage was only determined by the GC3s composition. As seen in Fig. 2, a majority of the points were not on the ENC curve. This implied that codon usage in *T. pisiformis* genes was affected by other factors in addition to compositional constraints.

The variation on the first and the second dimensions explained 7.16% and 4.40% of the total codon variation, respectively. Fig. 3 shows the position of each reconstructed genes on the plane defined by the first and second principal axes generated by COA on RSCU values of reconstructed genes. The distance between genes on the plot is a reflection of their diversity in RSCU, with respect to the first two axes. To investigate the effect of the GC content

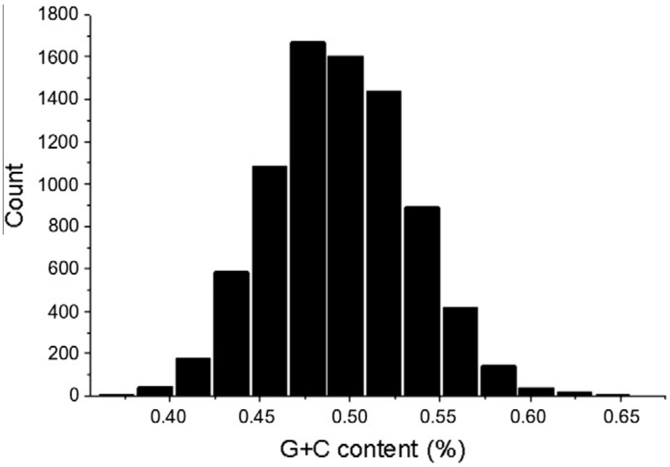

Fig. 1. Distribution of reconstructed genes GC contents in *T. pisiformis*.

Table 1
Mean values and standard deviation of GC, GC3s, and ENC values for reconstructed genes in *T. pisiformis*.

	N	Codons	GC3s (%)	GC (%)	ENC
Ribosome protein genes	59	9866	52.69 ± 6.87	50.30 ± 3.28	51.49 ± 6.45
Other protein genes	8059	802814	49.82 ± 7.13	49.47 ± 3.90	55.82 ± 4.41
All genes	8118	812680	49.84 ± 7.13	49.48 ± 3.90	55.78 ± 4.45

Table 2
Codon usage of *T. pisiformis* genes; AA: amino acid; *N*: the number of codons. The preferentially used codons are displayed in bold.

AA	Codon	*N*	RSCU	AA	Codon	*N*	RSCU
Ala	**GCU**	45524	1.30	Leu	**UUG**	33937	1.12
	GCC	38284	1.09		**CUU**	39632	1.30
	GCA	31761	0.91		**CUC**	41428	1.36
	GCG	24715	0.70		CUA	17904	0.59
Arg	**CGU**	26854	1.42		**CUG**	36893	1.21
	CGC	24579	1.30	Lys	AAA	41472	0.88
	CGA	22254	1.18		**AAG**	52821	1.12
	CGG	13883	0.74	Phe	UUU	33874	0.93
	AGA	13184	0.70		**UUC**	39075	1.07
	AGG	12508	0.66	Pro	**CCU**	25765	1.12
Asn	**AAU**	38069	1.05		**CCC**	23072	1.01
	AAC	34636	0.95		**CCA**	26002	1.13
Asp	**GAU**	54522	1.10		CCG	16969	0.74
	GAC	44541	0.90	Ser	**UCU**	26049	1.12
Cys	UGU	17870	0.96		**UCC**	27379	1.18
	UGC	19338	1.04		UCA	23335	1.00
Gln	CAA	32614	0.92		UCG	21063	0.90
	CAG	38283	1.08		AGU	21586	0.93
Glu	GAA	56449	0.93		AGC	20367	0.87
	GAG	65410	1.07	Thr	**ACU**	29031	1.17
Gly	**GGU**	36512	1.36		**ACC**	27437	1.11
	GGC	30139	1.12		ACA	23542	0.95
	GGA	27580	1.03		ACG	18939	0.77
	GGG	13277	0.49	Tyr	UAU	21083	0.81
His	CAU	21227	0.97		**UAC**	30723	1.19
	CAC	22602	1.03	Val	**GUU**	35811	1.19
Ile	**AUU**	41232	1.32		GUC	29881	0.99
	AUC	36117	1.16		GUA	16165	0.54
	AUA	16259	0.52		**GUG**	38575	1.28
Leu	UUA	12571	0.41				

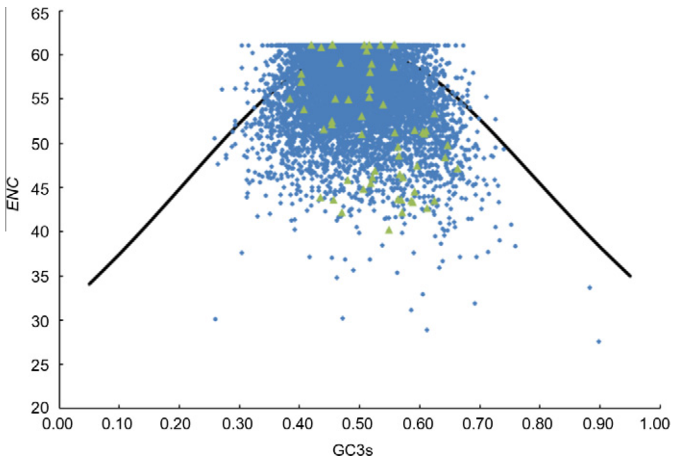

Fig. 2. ENC versus GC3s plot (ENC denotes the effective number of codons of each gene, and GC3s denotes the GC content on the third synonymous codon position of each gene) of reconstructed *T. pisiformis* genes. Green triangles and blue dots indicate ribosomal genes and other genes, respectively. The solid line represents the expected curve between GC3s and ENC under random codon usage. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

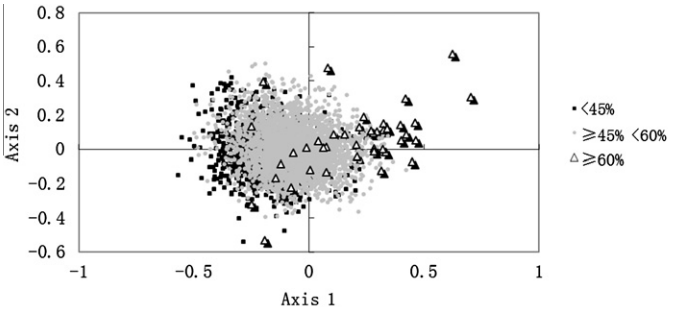

Fig. 3. Distribution of *T. pisiformis* reconstructed genes on the plane corresponds to the coordinates on the first and second axes produced by the correspondence analysis on RSCU. Black triangles, gray dots, and black dots indicate genes with a GC content higher than or equal to 60%, more than or equal to 45%, but less than 60% and less than 45%, respectively.

of genes on codon usage bias, different GC contents of genes were classified. Genes with a GC ⩾60% were plotted as black triangles, while genes with a GC <45% were plotted as black dots. Gray dots indicate the genes with GC content between 45% and 60%. This graph shows that the majority of genes near the origins of the axes clustered together to form an ellipsoid cloud in a range of −0.8 to +0.7 for the first axis and −0.6 to +0.6 for the second axis. The genes with a GC content <45% were mainly located on the left side of the plot, while most of the genes with a GC content between 45% and 60% were located in the middle of the plot (Fig. 3).

In order to analyze the codon usage of different kinds of gene, we selected the hydrophobic genes with gene scores >0.3, essential genes and ribosomal genes (i.e., highly expressed genes) from the 8118 genes. We could then redraw the position of each of the reconstructed genes on the plane defined by the first and second principal axes generated by COA according to the codon usage values of reconstructed genes. According to the data shown in Fig. 4, the analysis showed the most ribosomal genes and essential genes were clustered in the middle of the first major axis. The reconstructed genes with a Gravy score >0.3 were on the right side of the first major axis.

The gene positions on axis first were strongly positively correlated with the GC3s and GC content of individual *T. pisiformis* genes ($r = 0.825$ and 0.667, respectively, $P < 0.01$). These findings indicated that the genes, with higher GC3s and GC content values, were located on the right side of the first axis. Furthermore, there was a significant negative correlation between ENC and GC3s ($r = -0.165$, $P < 0.01$). From these results, we found that genes with lower GC3s and GC content values and higher ENC values had a lower codon bias. This showed that the variations in codon usage correlated with the nucleotide composition of the genes.

4.2. Analysis of fourfold degenerate codon families

If the unequal codon usage is due to biases in mutation patterns, then the expectation would be that the magnitude and the direction of the bias will be more or less the same for all codon families and for all genes, regardless of their function or expression levels. The nucleotide frequencies at fourfold degenerate codon families in the codon's third position are shown in Table 3. The results of

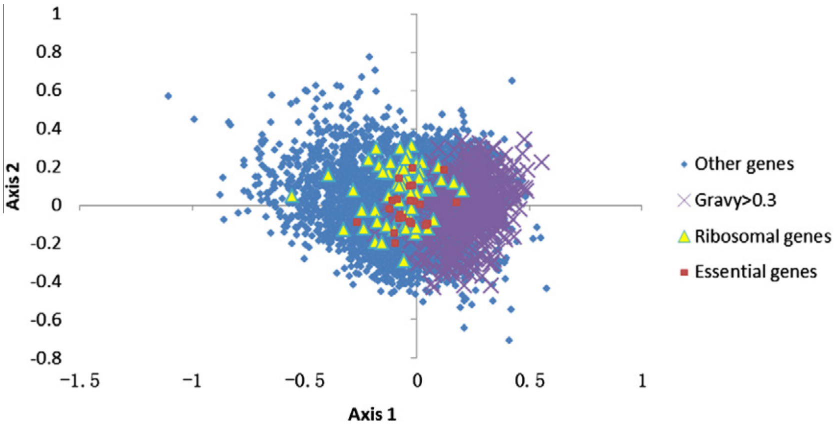

Fig. 4. Distribution of *T. pisiformis* reconstructed genes on the plane corresponds to the coordinates on the first and second axes produced by the correspondence analysis on codon usage. Red dots, yellow triangles, purple cross and blue dots indicate essential genes, ribosomal genes, genes with a Gravy value higher than 0.3 and other genes and other genes, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Nucleotide frequencies at fourfold degenerate codon families in the codon third position.

Amino acid (codon family)	T/A at the third position (%)	G/C at the third position (%)	AT/GC
Ser(UC)	50.5	49.5	1.02
Pro(CC)	56.4	43.9	1.29
Thr(AC)	53.1	46.9	1.13
Ala(GC)	55.1	44.9	1.23
Arg(CG)	56.1	43.9	1.28
Gly(GG)	59.6	40.4	1.48
Leu(CU)	42.4	57.6	0.73
Val(GU)	43.2	56.8	0.76

our analysis show that most of the T/A frequencies in fourfold degenerate codon families at the third position were more than the GC frequencies, except in the Leu and Val codon family. These results suggested that not only mutation pressure but also other factors influence the biases of synonymous codon usage.

4.3. Effect of gene expression level on synonymous codon usage bias

To explore the correlation between the codon usage bias and the gene expression level, we calculated the correlation coefficient between CAI against nucleotide composition and ENC. The set of reference sequences used to calculate CAI values in this study were genes that encode ribosomal proteins. It was found that there was one significantly negative correlation between the gene expression level assessed by the CAI and ENC values ($r = -0.329$, $P < 0.01$), and two significantly positive correlations between the CAI value with GC3s and GC content ($r = 0.783$ and 0.593, respectively, $P < 0.01$). All the above correlations were statistically significant, and we therefore conclude that codon usage in *T. pisiformis* was affected by the gene expression level. Taken together, the data suggested that genes with higher expression levels exhibited a greater degree of codon usage bias, were GC-rich and preferred codons with G or C at the synonymous position.

4.4. Effect of the hydrophobicity and aromaticity of encoded protein on synonymous codon usage bias

In order to investigate whether other factors in *T. pisiformis* genes could explain their codon usage, we performed a correlation analysis to evaluate whether Gravy and Aromo values were related to GC3s and ENC values. The correlation analyses between the hydrophobicity of each protein and GC3s and ENC values showed that the correlation coefficients ($r = 0.093$ and 0.028, respectively, $P < 0.05$) all correlated. The aromaticity of each protein also significantly correlated with GC3s and ENC values ($r = 0.058$ and 0.042, respectively, $P < 0.01$). The analysis results indicated that the degree of hydrophobicity and the aromatic amino acids were associated with the codon usage variation.

4.5. Translational optimal codons

The average RSCU values of high/low expressed gene sample group are listed in Table 4. Twenty-four codons, including UUC, CUC and CUG, were identified whose usage was significantly more frequent among the highly expressed genes ($P < 0.01$). All translational optimal codons (except CGU) ended on G or C.

5. Discussion

A transcriptome dataset was obtained from the deep sequencing of *T. pisiformis*. In this study, this dataset was applied to analyze the codon usage patterns of *T. pisiformis*. A total of 8118 reconstructed genes from 72,957 non-redundant consensus sequences were selected. When aligned with clean reads of *T. pisiformis*, more than 99.79% unigenes mapped to clean reads. CDSs with gaps were excluded, and only CDS sequences ⩾450 bp were used. The FrameDP program was used to analyze the CDSs. FrameDP is based on FrameD, which can identify ORFs by using extended interpolated Markov models (IMMs) and has frame shift correction ability. Unlike FrameD, FrameDP can use BLASTx results to generate training sequences and then calculate training matrices, which are expected to automatically represent the coding style of the species. The putative protein-coding sequences of *T. pisiformis* were generated based on their similarity with known proteins and on coding style recognition. The quality of sequences was effectively improved.

The exact number of genes in *T. pisiformis* is still unknown. Tapeworm genomes are small in size at approximately 110 Mb, compared to *Schistosoma* and *Macrostomum* genomes [13]. Hodgkin reported that the number of genes predicted for the *C. elegans* genome was approximately 20,000 [14]. Our reconstruction of 8118 genes should form a representative sample of *T. pisiformis*.

The third position of a codon is considered the most likely position to reflect genome base composition. The GC3s contents are different in different organisms. AT-rich organisms, such as

Table 4
Comparison of codon usage frequencies between highly and lowly expressed sequences of *T. pisiformis* genes. AA: amino acid; *N*: number of codons. Codon usage was compared using a Chi-squared contingency test to identify optimal codons. Asterisks denote codons that occurred significantly more often ($P < 0.01$).

AA	Codon	High	Low	AA	Codon	High	Low
		RSCU (*N*)	RSCU (*N*)			RSCU (*N*)	RSCU (*N*)
Phe	UUU	0.67 (1165)	1.14 (1377)	Ser	UCU	0.85 (769)	1.31 (1107)
	UUC*	1.33 (2302)	0.86 (1045)		UCC*	1.73 (1570)	0.71 (602)
Leu	UUA	0.18 (248)	0.84 (807)		UCA	0.73 (664)	1.39 (1170)
	UUG	0.84 (1138)	1.39 (1337)		UCG*	0.99 (894)	0.79 (667)
	CUU	0.95 (1287)	1.43 (1375)		AGU	0.72 (656)	0.98 (826)
	CUC*	2.07 (2791)	0.72 (693)		AGC*	0.97 (877)	0.82 (688)
	CUA	0.45 (608)	0.74 (711)	Pro	CCU	0.89 (824)	1.32 (1063)
	CUG*	1.51 (2036)	0.89 (857)		CCC*	1.53 (1421)	0.53 (424)
Ile	AUU	1.02 (1430)	1.39 (1558)		CCA	0.81 (755)	1.48 (1186)
	AUC*	1.66 (2318)	0.79 (886)		CCG	0.77 (711)	0.67 (539)
	AUA	0.31 (438)	0.81 (910)	Thr	ACU	1.01 (1087)	1.25 (1082)
Val	GUU	0.73 (1022)	1.61 (1613)		ACC*	1.63 (1758)	0.72 (623)
	GUC*	1.21 (1692)	0.72 (719)		ACA	0.64 (684)	1.26 (1097)
	GUA	0.39 (545)	0.74 (739)		ACG	0.72 (778)	0.77 (667)
	GUG*	1.66 (2315)	0.94 (941)	Ala	GCU	1.07 (1703)	1.47 (1745)
Tyr	UAU	0.53 (640)	1.09 (1063)		GCC*	1.47 (2344)	0.70 (835)
	UAC*	1.47 (1772)	0.91 (894)		GCA	0.69 (1102)	1.19 (1408)
His	CAU	0.64 (646)	1.28 (902)		GCG*	0.77 (1234)	0.63 (752)
	CAC*	1.36 (1371)	0.72 (510)	Cys	UGU	0.80 (714)	1.09 (697)
Gln	CAA	0.70 (1038)	1.08 (1440)		UGC*	1.20 (1078)	0.91 (582)
	CAG*	1.30 (1926)	0.92 (1238)	Arg	CGU*	1.50 (1210)	1.00 (639)
Asn	AAU	0.78 (1155)	1.20 (1708)		CGC*	2.27 (1835)	0.54 (344)
	AAC*	1.22 (1814)	0.80 (1132)		CGA	0.97 (779)	1.17 (752)
Lys	AAA	0.62 (1183)	1.05 (2268)		CGG	0.72 (584)	0.62 (397)
	AAG*	1.38 (2612)	0.95 (2061)		AGA	0.25 (199)	1.61 (1031)
Asp	GAU	0.87 (1774)	1.22 (2315)		AGG	0.29 (235)	1.07 (683)
	GAC*	1.13 (2312)	0.78 (1481)	Gly	GGU	1.30 (1538)	1.20 (1142)
Glu	GAA	0.65 (1573)	1.16 (2972)		GGC*	1.54 (1818)	0.72 (688)
	GAG*	1.35 (3295)	0.84 (2132)		GGA	0.78 (925)	1.52 (1451)
					GGG	0.38 (449)	0.56 (530)

Onchocerca volvulus, Mycoplasma capricolum and *Plasmodium falciparum*, tend to have poor C usage and commonly use A or T in the third position [15]. Other organisms show a preference for G or C in the third position, such as *Triticum aestivum, Hordium vulgare* and *Oryza sativa* [16]. Our results suggest that the mean GC3s (49.84%) content in *T. pisiformis* is almost equal to the AT content. The mean GC content in this study was slightly different from other *Taenia* species that had a mean GC content of 51%. This might indicate sequence differences between *T. pisiformis* and other *Taenia* species.

ENC values of each gene were plotted against its corresponding GC3s. The curve shows the expected position of genes whose codon usage was only determined by variation in GC3s content. If a particular gene is subject to GC compositional constraints, it will lie on or just below the expected curve. If a gene is subject to selection for translationally optimal codons, it will lie considerably below the expected curve. The *T. pisiformis* genes had a moderately high GC content and ENC values. A large number of points located near the solid curve of this distribution, suggesting that the codon usage of most of *T. pisiformis* genes are subject to other factors. The ENC values of each gene significantly correlated with its GC3s content, GC content, Gravy values and Aromo values. This indicated that the nucleotide composition, the degree of hydrophobicity and the aromatic amino acids are the factors that shape synonymous codon usage bias in *T. pisiformis*.

Acknowledgments

This work was supported by grants from the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) (No. IRT0848). We thank Hanmei Liu for their help with statistical analysis.

References

[1] R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pave, Codon catalog usage and the genome hypothesis, Nucleic Acids Res. 8 (1) (1980) R49–R62.

[2] A.T. Lloyd, P.M. Sharp, Evolution of codon usage patterns: the extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*, Nucleic Acids Res. 20 (20) (1992) 5289–5295.

[3] P.M. Sharp, W.H. Li, An evolutionary perspective on synonymous codon usage in unicellular organisms, J. Mol. Evol. 24 (1–2) (1986) 28–38.

[4] J.R. Powell, E.N. Moriyama, Evolution of codon usage bias in Drosophila, Proc. Natl. Acad. Sci. USA 94 (15) (1997) 7784–7790.

[5] Y. Zheng, W.M. Zhao, H. Wang, Y.B. Zhou, Y. Luan, M. Qi, Y.Z. Cheng, W. Tang, J. Liu, H. Yu, X.P. Yu, Y.J. Fan, X. Yang, Codon usage bias in Chlamydia trachomatis and the effect of codon modification in the MOMP gene on immune responses to vaccination, Biochem. Cell Biol. 85 (2) (2007) 218–226.

[6] T. Zhou, W. Gu, J. Ma, X. Sun, Z. Lu, Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses, Biosystems 81 (1) (2005) 77–86.

[7] A.A. Salamov, V.V. Solovyev, Ab initio gene finding in Drosophila genomic DNA, Genome Res. 10 (4) (2000) 516–522.

[8] K. Lin, Y. Kuang, J.S. Joseph, P.R. Kolatkar, Conserved codon composition of ribosomal protein coding genes in Escherichia coli, Mycobacterium tuberculosis and Saccharomyces cerevisiae: lessons from supervised machine learning in functional genomics, Nucleic Acids Res. 30 (11) (2002) 2599–2607.

[9] J.W. Fickett, Recognition of protein coding regions in DNA sequences, Nucleic Acids Res. 10 (17) (1982) 5303–5318.

[10] D. Yang, Y. Fu, X. Wu, Y. Xie, H. Nie, L. Chen, X. Nong, X. Gu, S. Wang, X. Peng, N. Yan, R. Zhang, W. Zheng, G. Yang, Annotation of the transcriptome from Taenia pisiformis and its comparative analysis with three Taeniidae species, PLoS one 7 (4) (2012) E32283.

[11] R. Zhang, H.Y. Ou, C.T. Zhang, DEG: a database of essential genes, Nucleic Acids Res. 32 (2004) D271–D272.

[12] F. Wright, The ‘effective number of codons’ used in a gene, Gene 87 (1) (1990) 23–29.

[13] P.D. Olson, M. Zarowiecki, F. Kiss, K. Brehm, Cestode genomics–progress and prospects for advancing basic and applied aspects of flatworm biology, Parasite Immunol. 34 (2–3) (2012) 130–150.

[14] J. Hodgkin, What does a worm want with 20,000 genes, Genome Biol. 2 (11) (2001) 2001–2008.

[15] J.G. Waterkeyn, C. Gauci, A.F. Cowman, M.W. Lightowlers, Codon usage in Taenia species, Exp. Parasitol. 88 (1) (1998) 76–78.

[16] A. Kawabe, N.T. Miyashita, Patterns of codon usage bias in three dicot and four monocot plant species, Genes Genet. Syst. 78 (5) (2003) 343–352.