Comprehensive analysis of codon usage pattern of whole genome in protozoa, *Stylonychia lemnae*

ying wang (✉ yingyingWY2018@outlook.com)

Harbin Normal University <https://orcid.org/0000-0001-9263-1034>

Lin Yao

Harbin Normal University

Jinfeng Fan

Harbin Normal University

Xueying Zhang

Harbin Normal University

Changhong Guo

Harbin Normal University

Ying Chen

Harbin Normal University

Research article

Keywords: *S. lemnae*, macronucleus, codon usage, mutation pressure, natural selection

Posted Date: November 15th, 2019

DOI: <https://doi.org/10.21203/rs.2.16499/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Codon usage pattern is an important evolutionary feature in genomes widely observed in many organisms. *Stylonychia lemnae* is a classical model single-celled eukaryote, and a quintessential ciliate typified by dimorphic nuclei: a germline micronucleus and a vegetative macronucleus. Analysis of codon usage pattern of *S. lemnae* macronucleus genome helps in understanding evolution at molecular level and acquires significance in mRNA translation, design of transgenic and new gene discovery.

Results: The codons of the macronucleus genome sequence of *S. lemnae* were analyzed and 20,750 coding sequences (CDS) were screened. The overall codon usage of *S. lemnae* is similar and slightly biased. The value of effective number of codons (ENC) showed that the overall extent of codon usage bias in *S. lemnae* is relatively high. Nucleotide analysis showed that the overall codon usage is biased toward A- and U-ending codons. The phylogenetic analysis indicated that ciliate is independent evolutionary origins from a common ancestor. The RSCU analysis showed that the codon usage pattern of *S. lemnae* is more similar to that of *Thrahymana thermophila* and *Paramecium caudatum*. Correlation analysis, ENC-GC 3S plot, and PR2 plot indicated that the codon usage patterns of *S. lemnae* are influenced by both mutational pressure and natural selection, neutrality plot analysis showed that those two factors play major roles.

Conclusions : Codon usage patterns in eukaryotes are not determined by translational efficiency, but also are determined by the genome. Our study is the first attempt to evaluate the codon usage pattern of *S.lemnae* macronucleus genome to better understand the evolutionary changes. These results built the base for further research on the molecular evolution of *S. lemnae*.

Background

The ciliate *S. lemnae* is a classical model single-celled eukaryote, which widespread in ponds, rivers and marshes. *S. lemnae* possesses both a macronucleus (MAC), specialized for gene expression, and a micronucleus (MIC), containing the germline genome that permits recombination and transmission of genetic information across sexual generations [1]. The macronucleus genome of *S. lemnae* is 50.2-Mb, contains 80 ribosomal relate genes [2].

It is well known that most of the amino acids (except Methionine and Tryptophan) are coded by more than one codon, which due to the degeneracy of the genetic codes. And this use of more than one codon for the same amino acid is called synonymous codon usage[3]. Furthermore, the synonymous codons usage termed as codon usage bias, which means synonymous codons usage is not random and for the same amino acids, some codons might be preferred over others[4].

Ciliate is a complex protozoan with two nuclei and its codon usage bias is not universal [5, 6]. With the exception of a few species, most of metazoans use the standard genetic coding system for protein translation, as do viral and prokaryote. However, in single-celled organisms, the genetic code expresses its particularity. *Stylonychia* are hypotrichous ciliates with alternative genetic codes. It has reassigned the

canonical stop codons TAR to glutamine codons, resulting in genomes with only one stop codon (TGA) [7,8,9]. Whereas in *Oxytricha*, UAA and UAG encode glutamine, TGA is only one stop codon. Furthermore, Non-standard glutamine codons are not used in the four species of Euplotes [10].

The codon usage patterns in *S. lemnae* was analyzed by Martindale D.W in 1989[11], but only two tubulin protein-coding gene sequences were available for analysis. Codon usage has been shown to affect protein structure and function through interfering with translation kinetics, and cotranslational protein folding [12]. Furthermore, synonymous codon usage patterns can be an essential tool in revealing specie evolution [13]. The genome data were deficient for *S. lemnae* codon analysis in previous studies, and knowledge on the evolution of DNA in *S. lemnae* on the other hand is more scarce. So it is necessary to clear the influencing factor on its codon usage with each other, such as mutation pressure [14] or natural selection [15], in order to fill the large existing gaps of knowledge and data.

Heritable Variations in *Stylonychia sp* is reported firstly in 1915, and causing who the main fact has considered to be cumulative effects of selection [16]. But due to the lack of genomic data, the molecular mechanism of genetic change can not be elucidated. Basing on the reason, we have downloaded all the complete *S. lemnae* macronuclear genome coding sequences of until June the 21st, 2019 and have analyzed their codon usage patterns.

Results

Phylogenetic analysis of S. lemnae based on 18s ribosomal DNA (18s-rDNA)

To determine the phylogenetic relationship of *S. lemnae*, protozoan and metazoans, a phylogenetic tree was drawn (Fig.1). The results show that *S. lemnae* clusters with *Stylonychia sp.* with high support (ML/BI:86/0.76), which is then sister to *Tetrahymena sp.* and *S. cerevisiae* with low bootstrap value (ML/BI: 78/0.76, ML/BI: 60/0.45 respectively); which clusters with *C.elegans* and *H.sapiens* in a poorly supported clade (ML/BI: 55/0.43); and also groups with *Paramecium sp.*...

Codon composition analysis

CodonW software was used to analyze the macronucleus genome coding sequence of the *S. lemnae*. The GC content ranged from 22.1% to 58.6% in the gene of *S. lemnae*. The GC content was mainly distributed in 25% ~ 40% with an average of 33.7%, indicating that AT was enriched in the genome (Supplementary Fig. S1).

Codon usage pattern

To investigate the codon usage bias on *S. lemnae* codon usage pattern, the RSCU values in coding sequences were assessed. Among 26/ preferred codons of corresponding all amino acids (except Methionine and Tryptophan) in *S. lemnae* coding sequences, 26 are A/U-ended (twelve A-ended; fourteen U-ended) and the remaining are C/G ended (one C-ended; one G-ended). Therefore, most preferred codons in *S. lemnae* are found to be A-ended or U-ended codons (Table

1). By analyzing the 26 preferred codons, we can find that the RSCU values of six codons, GGA (G), UCA (S), CCA(P), ACU(T), GCU (A) and AGA(R) are >1.6, whereas the RSCU values of the remaining are also found to be > 1 and < 1.6 (Table 1). Nucleotide composition (A/T-rich) and RSCU analysis (A/U-ended) show that the preferred codons have been mostly influenced by compositional constraints of nucleotide, which account for mutation pressure. Additionally, it was found 25 are A/U-ended in highly expressed genes in *S. lemnae*, but 21 are C/G end in lowly-expressed genes (Table 2), which also show A-ended or U-ended codons are preferentially codons of highly expressed genes. These results indicate that mutation pressure decide the third position of codons and affect codon usage pattern.

To investigate the potential influences (mutation pressure or natural selection) of biological evolution on the codon usage patterns, the RSCU values of the codons in *S. lemnae* coding sequences were calculated and then were compared with five model organism (*H. sapiens*, *C. elegans*, *S. cerevisiae*, *T. thermophila* and *P. caudatum*).. We find that preferred codons is eight between *S. lemnae* and *H. Sapiens*; fourteen between *S. lemnae* and *C. elegans*; twenty one between *S. lemnae* and *S. cerevisiae*; twenty one between *S. lemnae* and *T. thermophila*; and nineteen between *S. lemnae* and *P. caudatum* (Table 1). Those results show the similarity in codon pattern between *S. lemnae* and *H. Sapiens* is lower than that among *C. elegans*, *S. cerevisiae*, *T. thermophila* or *P. caudatum*. The codon usage bias of *S. lemnae* differs greatly from that of higher eukaryotes and is similar to that of lower eukaryotes. It is generally believed that the closer the species are, the more similar the codon usage pattern should be.

Mutational pressure versus natural selection on codon usage pattern

In order to get an insight whether codon usage bias of *S. lemnae* was driven by mutational pressure alone or other environmental factors, we analyzed correlation coefficients between overall nucleotide composition (A%, T%, G%, C%, GC%), nucleotide composition at the third position of codons (A3%, T3%, G3%, C3%, GC3%) and ENC (Table 3). The results show that the A content has a significant positive correlation with the content of A3s and G3s, but has a significant negative correlation with the content of C, T, G, GC, C3s, T3s and GC3s. The C content has a significant positive correlation with the content of G, GC, C3s, GC3s and ENC, but has a significant negative correlation with the content of A, T, A3s T3s and G3s. The T content has a significant positive correlation with T3s contents, but has a significant negative correlation with the content of A, C, G, GC, A3s, C3s, G3S, GC3s and ENC. The G content has a significant positive correlation with the content of G, GC, C3s, G3s, GC3s and ENC, but has a significant negative correlation with A, T, A3s and T3s content. The GC contents has a significant positive correlation with the content of CG, G3s, C3s, GC3s, but has a significant negative correlation with the content of A, T, A3s and T3s. The ENC value has a significant positive correlation with the content of C, G, GC, C3s, G3s and GC3s, but has a significant negative correlation with the content of T, A3s and T3s, and has no correlation with the content of A. These results suggested that mutation pressure may affect the codon usage bias for *S. lemnae*.

To study the relative contribution of two major factors, i.e., natural selection and mutational pressure on codon usage pattern, we performed PCA analysis taking RSCU scores to find out major trends of codon

usage in *S. lemnae* genes. A plot of PC1 and PC2 showed important features of the codon usage pattern in *S. lemnae* genes (Fig. 2a). From this analysis major trends in codon usages were detected in which axis1 (PC1) accounted for 13.4%, whereas axis2 (PC2), axis3 and axis4 accounted for 10.2%, 6.6%, and 4.6% of total variation in *S. lemnae*. Axis1-axis4 explaining 34.8% of the cumulative variances, which indicates that no single factor influence codon usage pattern in *S. lemnae*.

In order to characterize the codon usage patterns from different types of genes, ribosomal related genes was statistical analysis by PCA (Fig. 2b). It was clearly seen that ribosomal genes of *S. lemnae* were clustered on the right side of PC1, and indicate that compositional constraints are a major factor in CUB, that is mutation pressure mostly shaped its codon pattern, but other factors are also powerful.

Additionally, Correlation analysis was also performed to determine the correlations between the first two axes and nucleotide constraints of *S. lemnae* genome (Table 4). The results show that the Axis1 is positively correlated with the A and A3s, whereas it is negatively correlated with the contents of C, G, GC, C3s, GC3s and ENC. Meanwhile, Axis2 is insignificantly correlated with the C, T, G, GC, A3s, C3s, T3s, G3s, GC3s and ENC. Moreover, these results also support the key role of mutational pressure in shaping the codon usage bias in *S. lemnae*.

To investigate the effect of natural selection on codon usage bias in *S. lemnae*, correlation analysis between the characters of amino acid (Gravy values and Aromaticity) and the codon bias (Axis1, Axis2, ENC, and GC 3s) was performed (Table 5). The result indicates that Axis1 has a significant negative correlation with AROMA and GRAVY, and Axis 2 has a significant positive correlation with AROMA and GRAVY, but Axis1 or Axis 2 has non-significant correlation with ENC, GC or GC 3s. Earlier it is found that AROMA and hydrophobicity of the encoded proteins has a significant correlation with the base composition of third codon positions in some other prokaryotes, several eukaryotes and viral genomes [17,18,19,20,21]. However, there is no report of such correlation in any of the ciliate genomes studied so far. To our knowledge, this is for the first time, the aromaticity and hydrophobicity of amino acid under natural selection have been demonstrated important effect on the codon usage pattern of *S. lemnae*.

ENC- GC 3S plot analysis

To study the factors (mutation pressure, natural selection) influencing the codon usage patterns of *S. lemnae* coding sequences, we performed ENC- plot analysis for GC 3S, PR2 plot and neutrality plot analysis.

The degree of codon bias is reflected by the size of ENC value. ENC value ranges from 20 to 61, with the level of base composition bias increasing as the ENC values approach 20. Similarly, genes expressed at low levels contain numerous rare codons, with a higher ENC value approach 61. The convention uses 35 as the criterion for biased bias [22]. The ENC values of the *S. lemnae* genome range from 24.8 to 61, and most of them are more than 35, so the codon bias of the gene is weak (Supplementary Table S2). The average GC3 content is 33.7%, GC1 and GC2 are 38.68% and 30.1%, respectively, indicating that the codon base composition is mostly A and U.

The association analysis between ENC - GC3 is shown in Fig. 3. If ENC value of genes will lie on or just below the continuous curve of the expected ENC values, it indicates that the codon bias is only constrained by a G3+ C3 mutational bias [23]. In the figure, more ENC value of genes loci are on the top or far below the curve of the expected ENC values, but a little ENC value of genes lie on or just below the expected curve. It indicates that the codon usage patterns have been influenced by mutation pressure and other factors, including natural selection.

PR2-plot plot analysis.

The relationship between purine(A and G) and pyrimidines (T and C) of partial amino acids of each gene was analyzed by PR2-plot mapping. There is no bias in the selection or mutation pressure when the plot lie on the center, where both coordinates are 0.5[24]. According to Supplementary Fig. S2, most of the genes are distributed on the high right of the plan, indicating that the frequency of A is higher than T, and the frequency of G is higher than C. those results demonstrate that the codon usage pattern of *S. lemnae* is also shaped by mutation pressure and other factors, including natural selection.

Neutral Plot Analysis

Neutrality plot analysis was used to estimated the relationship between GC12 and GC3, and it elucidates the extent of influence of mutation-selection equilibrium in shaping the CUB. Synonymous mutation usually occurs at the third base of codons. If the correlation between GC12and GC3 is statistically significant and the slope of the regression line is close to 1, it is assumed that mutation pressure is the main force shaping the codon usage bias of genes. But selection against mutation bias is the dominant factor, the slope of the regression line is close to 0 with no correlation between GC12 and GC3. The analysis show that no correlation is observed between the value of GC 12 and GC 3 ($r = 0.286$, $P > 0.05$) which seemed indicative of natural selection playing a little role in codon usage bias of *S. lemnae* genome (Fig. 4). Then, after calculating the slope of the regression in the neutrality plot, this was the case. The slope of the regression line was calculated to be 0.2016, high-lighting the relative GC 3 (natural selection) is 79.94%, while the relative constraint on neutrality (mutation pressure) is 20.16%. These results suggested that natural selection is the dominant factor while mutation pressure played a minor role in shaping codon usage bias of *S. lemnae* genes.

Conclusion And Discussion

According to our result, Phylogenetic relationships of *S. lemnae* are similar to *Tetrahymena sp.*, *Paramecium sp.* and *S. cerevisiae*, but are far from *C. elegans* and *H. sapien*. It indicated that [the origin and early evolution of many eukaryotes](#) remain unknown or ambiguous, the codon usage pattern is non-universal and influenced by evolutionary processes.

Codon usage of plays a significant role in the evolution of metazoans, whereas our understanding of that in the evolution of protozoan is limited. Until now, the codon usage pattern of protozoan genome, such as *Tetrahymena thermophila* and *Paramecium tetraureli*[25], had been only reported. In the present study,

comprehensive analyses of codon usage of *S. lemnae* based on macronuclear genome coding sequences were performed. We observed that (1) *S. lemnae* coding sequences were AU-rich;(2)A-and U-ending codons were preferred in *S. lemnae* macronuclear genome coding sequences;(3) a relatively stable genomic composition with a high codon usage bias was observed in *S. lemnae* coding sequences. (4) natural selection and mutation pressure shaped the codon usage in *S. lemnae* macronuclear genome coding sequences.

Research reported that the overall nucleotide composition can influence codon usage bias[26]. Our nucleotide composition results showed that the *S. lemnae* macronuclear genome were AU-rich. The organisms with AT-rich genome, such as *T. thermophila* and *P. tetraurelia*, tend to use A or T at the third position in coding sequence [27,28,29]. Likewise, AT-rich genome in *S. lemnae* prefers A/T as the third codon. whereas GC-rich genomes tend to contain codons ending with G and C, such as Archaea, Bacteria, *Hordeum vulgare*, *Triticum aestivum*, fungi and *Oryza sativa* [30,31,32]. This trend supports the influence of mutation pressure.

However, codons end in highly expressed genes present a very different pattern in *Tetrahymena*, *Paramecium* and *S. lemnae*. Most optimal codons end in T/A, rather than G/ C in RSCU analysis ($T > A > C > G$) in *S. lemnae*. But C (and to a lesser extent T) end is preferred in highly expressed genes in *Tetrahymena* and *Paramecium*, and A ending (in *Tetrahymena*) and G ending (in *Paramecium*) tend to be avoided[25]. Research found the third position of codons in *Tetrahymena* and *Paramecium* is usually under less selective pressure than the other two positions[25], so mutational pressure is the main factor affecting the codon usage bias of protozoan.

We observed that the ENC values of the *S. lemnae* macronuclear genome coding sequences were lower than 35 (Tables S2), indicative of a high degree of preference. Normally, the high A/T content of *S. lemnae* coding region, and which also has a significant negative correlation with ENC, that may be the reason of low ENC value. Additionally, GC 3S and G 3S are strongly positively correlated with ENC, which indicate that the codon bias under mutation pressure is relatively high. PR2 plot also support that T and C are used more frequently than A and G in the third base of synonymous codons in *S. lemnae* complete coding regions, respectively. Likely, T/A ends of codons are also used more frequently than C/G ends in *Tetrahymena* and *Paramecium*[25]. In another respect, because the A, T, C, G, GC, A3s, T3S, C3S, G3s and GC3s content have a correlation with the two principle axes, clearly showed that codon usage pattern of *S. lemnae* macronuclear genome coding sequences from base composition is influenced by mutation pressure.

Additionally, ENC-plot, PR2 plot, neutrality plot and correlation analysis between GRAVY and aromaticity values and those of the ENC, GC3S, GC, and first two principle axes demonstrated that natural selection is also the major factor that has contributed to shaping codon usage of *S. lemnae* macronuclear genome coding sequences. It firstly indicates the role of hydropathicity and aromaticity forming the codon usage pattern of the clate.

In short, our analysis revealed that the codon usage bias in *S. lemnae* is high and natural selection such as aromaticity, hydropathicity, is the main factor that affects codon usage variation in *S. lemnae*. Mutation pressure of nucleotide composition is also an important factor influencing codon usage bias. The evolution of *S. lemnae* probably reflects a dynamic process of mutation and natural selection to adapt its codon usage. This study of codon usage patterns in *S. lemnae* is the first reveal information about molecular evolution, and also builds the base for analysis alternative genetic codes of hypotrichous ciliates.

Methods

Sequence data

The complete *S. lemnae* genome is obtained from the *S. lemnae* genome database (<http://stylo.ciliate.org/index.php/>). A total of 21,061 coding sequences were downloaded. Then we deleted the short sequence whose length is less than 300 bp. Finally, we extracted 20,750 CDSs which had correct start and stop codons with exact multiple of three bases to do data analysis.

Nucleotide composition analysis

The component parameters of *S. lemnae* macronucleus genomes were calculated after removing five non-bias(AUG,UGG,UAA,UGA,UAG),.The over all nucleotide composition(A,C,T and G%) and the total AU and GC content were calculated using BioEdit.The nucleotide compositions were calculated by CodonW software 1.4.2 package(<http://codonw.sourceforge.net>), including (i) frequency of occurrence of A, C, U, G and GC, respectively); (ii) frequency of occurrence of A,T,G and C at the third position of the synonymous codons, respectively; (iii) frequencies of occurrence of the GC at the 1st, 2nd and 3rd codon positions (GC1s, GC2s, GC3s); (iv) mean frequencies of GC at GC1s and GC2s

Relative Synonymous Codon Usage (RSCU) Analysis

The RSCU value of were defined as between the observed and expected usage frequency with the assumption that all codons for a particular amino acid are used equally[33]. An RSCU value close to 1.0 indicates that there is no bias for that codon, whereas codons with RSCU values >1.0 has a positive codon usage bias, and codons with RSCU values <1.0 has a relatively negative codon usage bias, then codons with RSCU values >1.6 are considered to be"over represented"[34]. In this study, the RSCU values of *S. lemnae* were calculated by CodonW software, the RSCU values for *H.sapiens* were retrieved from Singh R. K.[35], the RSCU values of *C.elegans* were retrieved from Stenico M [36],the RSCU values of *S.cerevisiae* were retrieved from Lloyd A.T [37], and the the RSCU values of *P.caudatum* and *T.thermophila* were retrieved from Barth D and Hannah M.W [38,39], respectively.

The RSCU values were calculated using the following equation:

$$RSCU_{ij} = \frac{g_{ij}}{\sum_j^{n_i} g_{ij}} \times n_i$$

Where g_{ij} is the observed number of the i -th codon for the j -th amino acid that has an " n_i " type of synonymous codons.

Effective number of codons analysis.

The ENC was calculated the absolute codon usage bias of of gene and genome level, which is the best estimator of absolute codon usage bias [40,41]. The ENC value is between 20(only one synonymous codon is used, an extreme codon usage bias) and 61 (the synonymous codons are used equally, no bias). When the ENC value ≤ 35 is generally considered that the gene has significant codon bias[42].

Determination of optimal codons

Optimal codons definition referred to the analytical method of Liu et al [43]. Gene expression levels were determined by CAI values. The method is as follows: Calculate the CAI of each gene, arrange it of each gene based on CAI values, and take 5% of total genes with the highest CAI valuea to form a high. The RSCU values of each codon in the two data sets is calculated, and then the RSCU values of the high dataset is subtracted that of the low dataset. If difference is greater than 0.08, the corresponding codon is defined as the optimal codon. The software used in this study includes CodonW 1.4.2, SPSS 20 and Excel 2016.

ENC-GC_{3S} plot analysis

The ENC- GC3S plot consists of plotting GC3s values in the abscissa and the ENC values in the ordinate, which was used to investigate the role of mutational pressure in codon usage bias. The expected ENC value for each GC 3S was calculated using the following formula: $ENC_{expected} = 2 + GC3 + 29/[GC_3^2 + (1 - GC_3)^2]$ [44].

Parity rule 2 analysis

parity rule 2 (PR2) mapping analysis ($A_3/(A_3+U_3)$ plotted against $G_3/(G_3+C_3)$) was to used investigating the effects of mutation and natural selection on codon usage. In this PR2 mapping, the value of a AU bias [$A_3/(A_3+U_3)$] at the third base of the codon is in the abscissa and the value of G C bias [$G_3/(G_3+C_3)$] is in the abscissa[45].

Neutral Plot Analysis

The influence of mutation bias and natural selection on codon usage were tested by Neutral maps by plotting GC12 values against GC3 values. The mutation pressure is represented by the slope of the regression line plotted between the GC12 and GC3 contents, where regression lines that fall near the diagonal (slope = 1.0) indicates no or weak external selection pressure. In contrast, if regression curves deviate from the diagonal, it indicates a significant influence of natural selection on codon usage bias[21,46].

Principal Component Analysis (PCA).

PCA was used to analyze the major trends in synonymous codon usage of genes mutation. We normalize the data according to the method by Sharp and Li [47]. In each PC, the score of the gth gene (yg) gene was normalized by the mean (m) and standard deviation (SD), expressed as $z_g = (y_g - m) / SD$. Highly expressed genes were a z_g score greater than 5.17 (theoretically covering only the range of 1.5% of total genes). Then gene expression levels were identified as the major trends in PC score changes in genes. In addition, we analyzed the distribution of PC scores for constitutively overexpressed genes (encoding ribosomal proteins).

Phylogenetic analysis

A phylogenetic tree was constructed based on the nucleotide sequences of the coding regions of *S. lemnae*, using maximum likelihood (ML) method with a bootstrap value of 1000 replicates in CIPRES Science Gateway using RAxML-HPC2 on XSEDE v8.2.10[48]. Bayesian inference (BI) analysis was carried out with MrBayes v3.2.6 on XSEDE [49] on the CIPRES Science Gateway.

List Of Abbreviations

CUB	Codon usage bias
CDS	coding DNA sequences
RSCU	Relative synonymous codon usage
ENC	effective number of codon
PCA	Principal component analysis
PC	Principal component
SD	standard deviation
CAI	Codon adaptation index
<i>S. lemnae</i>	<i>Stylonychia lemnae</i>
<i>H. sapiens</i>	<i>Homo sapiens</i>
<i>C. elegans</i>	<i>Caenorhabditis elegans</i>
<i>S. cerevisiae</i>	<i>Saccharomyces cerevisiae</i>
<i>T. thermophila</i>	<i>Tetrahymena thermophila</i>
<i>P. caudatum</i>	<i>Paramecium caudatum</i>

Declarations

Ethics approval and consent to participate.

Because the organisms studied in this paper are single-celled eukaryotes, they do not involve ethical issues.

Consent for publication

Not applicable” in this section.

Availability of data and materials.

Data source [address](http://stylo.ciliate.org/index.php/home/welcome/) <http://stylo.ciliate.org/index.php/home/welcome/>

Competing interests

The authors declare that they have no competing interests” in this section.

Funding.

The workstation used for data processing was financially supported by the National Natural Science Foundation of China (31501893, 31471950) and Natural Science Foundation of Heilongjiang province, China (QC2018021) .

Authors' contributions

YL analyzed and interpreted data and was a major contributor in writing the manuscript. WY and CY carried out previous data processing. JF, XZ and CG interpreted the data. All authors have read and approved the final manuscript.

Acknowledgement

Sequence data was Institute of aquatic biology, Chinese Academy of Sciences provides the transcriptome data of *Stylonychia lemnae*.

References

1. Prescott, D M. "The DNA of ciliated protozoa. " Microbiol Rev58.2(1994):233–267.
2. Mu, Weijie, et al. "Epidermal growth factor-induced stimulation of proliferation and gene expression changes in the hypotrichous ciliate, *Stylonychia lemnae*." GENE592.1(2016):186–192.
3. Mehmood, Butt Azeem, et al. "Genome-Wide Analysis of Codon Usage and Influencing Factors in Chikungunya Viruses." PLoS ONE 9.3(2014):e90905-.
4. Tourancheau, A. B., et al. "Genetic code deviations in the ciliates: evidence for multiple and independent events. " The EMBO Journal 14.13(1995):3262–3267.
5. Heaphy, Stephen M., et al. "Novel ciliate genetic code variants including the reassignment of all three stop codons to sense codons in *C. magnum*." Molecular Biology and Evolution(2016):msw166.
6. Chen, Lin, et al. "Analysis of codon usage patterns in *Taenia pisiformis* through annotated transcriptome data." Biochemical and Biophysical Research Communications 430.4(2013):1344–1348.
7. Helftenbein, E. "Nucleotide sequence of a macronuclear DNA molecule coding for alpha-tubulin from the ciliate *Stylonychia lemnae*. Special codon usage: TAA is not a translation termination codon. " Nucleic Acids Research 13.2(1985):415–33.
8. Lekomtsev, S., et al. "Different modes of stop codon restriction by the *Stylonychia* and *Paramecium* eRF1 translation termination factors." Proceedings of the National Academy of Sciences 104.26(2007):10824–10829.

9. Lozupone, Catherine A., R. D. Knight, and L. F. Landweber. "The molecular basis of nuclear genetic code change in ciliates." *Current Biology* 11.2(2001):65–74.
10. Tourancheau, A. B., et al. "Genetic code deviations in the ciliates: evidence for multiple and independent events. " *The EMBO Journal* 14.13(1995):3262–3267.
11. Martindale, D. W.. "CODON USAGE IN TETRAHYMENA AND OTHER CILIATES." *The Journal of protozoology* 36.1(1989):29–34.
12. Athey, John, et al. "A new and updated resource for codon usage tables." *BMC Bioinformatics* 18.1(2017):391.
13. Wong, Emily Hm, et al. "Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus." *Bmc Evolutionary Biology* 10.1(2010):253–0.
14. Sharp, P.M., et al. "DNA sequence evolution: the sounds of silence. " *Philosophical Transactions of the Royal Society of London* 349.1329(1995):241–247.
15. Reis, and M. D.. "Solving the riddle of codon usage preferences: a test for translational selection." *Nucleic Acids Research* 32.17(2004):5036–5044.
16. Middleton, A. R. "Heritable Variations and the Results of Selection in the Fission Rate of *Stylonychia pustulata*." *Proceedings of the National Academy of Sciences of the United States of America* 19.4(1915).
17. Singh, Niraj K., et al. "Characterization of codon usage pattern and influencing factors in Japanese Encephalitis Virus." *Virus Research* (2016):S016817021630288X.
18. D'Onofrio, G., Jabbari, K., Musto, H., & Bernardi, G. (1999). The correlation of protein hydropathy with the base composition of coding sequences. *Gene*, 238(1), 3–14.
19. Jabbari, Kamel, et al. "The correlation between GC3 and hydropathy in human genes." *Gene (Amsterdam)* 317.none(2003):0–140.
20. Das, Sabyasachi, S. Paul, and C. Dutta. "Synonymous codon usage in adenoviruses: Influence of mutation, selection and protein hydropathy." *Virus Research* 117.2(2006):0–236.
21. Jun Tao, Huipeng Yao..Comprehensive analysis of the codon usage patterns of polyprotein of Zika virus. *Progress in biophysics and molecular biology*..(2018):05.001.
22. Mochizuki, Kazufumi."High efficiency transformation of *Tetrahymena* using a codon-optimized neomycin resistance gene." *Gene* 425.1–2(2008):0–83.
23. Mehmood, Butt Azeem, et al. "Genome-Wide Analysis of Codon Usage and Influencing Factors in Chikungunya Viruses." *PLoS ONE* 9.3(2014):e90905.
24. Sueoka, and N. "Directional mutation pressure and neutral molecular evolution. " *Proceedings of the National Academy of Sciences* 85.8(1988):2653–2657.
25. Salim, Hannah M. W., K. L. Ring, and A. R. O. Cavalcanti. "Patterns of Codon Usage in two Ciliates that Reassign the Genetic Code: *Tetrahymena thermophila* and *Paramecium tetraurelia*." *Protist* 159.2(2008):283–298.

26. Jenkins, Gareth M, and E. C. Holmes. "The extent of codon usage bias in human RNA viruses and its evolutionary origin." *Virus Research* 92.1(2003):0–7.
27. Liu, and Xiong'en. "A more accurate relationship between 'effective number of codons' and GC3s under assumptions of no selection." *Computational Biology and Chemistry* 42.Complete(2013):35–39.
28. "Macronuclear Genome Sequence of the Ciliate *Tetrahymena thermophila*, a Model Eukaryote." *PLoS Biology* 4.9(2006):e286.
29. Aury, Jean Marc, et al. "Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*." *NATURE* 444.7116(2006):171–178.
30. Ruth, Hershberg, D. A. Petrov, and M. W. Nachman. "General Rules for Optimal Codon Choice." *PLoS Genetics* 5.7(2009):e1000556.
31. Kawabe, Akira, and N. T. Miyashita. "Patterns of codon usage bias in three dicot and four monocot plant species." *Genes & Genetic Systems* 78.5(2003):343–352.
32. Yang, Xing, X. Luo, and X. Cai. "Analysis of codon usage pattern in *Taenia saginata* based on a transcriptome dataset." *Parasites & Vectors* 7.1(2014):527.
33. Sharp, P.M.. "An evolutionary perspective on synonymous codon usage in unicellular organisms." *J. Mol. Evol.* 24(1986).
34. "Analysis of amino acid and codon usage in *Paramecium bursaria*." *Febs Letters* 589.20(2015):3113–3118.
35. Singh, R. K., and S. P. Pandey. "Phylogenetic and Evolutionary Analysis of Plant ARGONAUTES. " (2017).
36. Stenico, M., A. T. Lloyd, and P.M. Sharp. "Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases." *Nucleic acids research* 22.13(1994).
37. Lloyd, Andrew T., and P.M. Sharp. "Evolution of codon usage patterns: The extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*." *Nucleic Acids Research* 20.20(1992):5289–5295.
38. Barth, Dana, and T. U. Berendonk. "The mitochondrial genome sequence of the ciliate *Paramecium caudatum* reveals a shift in nucleotide composition and codon usage within the genus *Paramecium*." *BMC Genomics* 12.1(2011):272.
39. Salim, Hannah M. W., K. L. Ring, and A. R. O. Cavalcanti. "Patterns of Codon Usage in two Ciliates that Reassign the Genetic Code: *Tetrahymena thermophila* and *Paramecium tetraurelia*." *Protist* 159.2(2008):283–298.
40. Nasrullah, Izza, Butt, Azeem M., Tahir, Shifa, Idrees, Muhammad, Tong, Yigang,. Genomic analysis of codon usage shows influence of mutation pressure, natural selection, and host features on Marburg virus evolution. *BMC Evol. Biol.* (2015):15, 174.
41. Wright, F.. The 'effective number of codons' used in a gene. *Gene*,(1990):87(1), 23–29.

42. Wang, H, et al. "Correction: Analysis of Synonymous Codon Usage Bias of Zika Virus and Its Adaption to the Hosts." Plos One 11.11(2016):e0166260.
43. Booth, L., Wolfe, B., & Doerder, F. P. Molecular polymorphism in the mta and mtb mating type genes of tetrahymena thermophila and related asexual species. Journal of Eukaryotic Microbiology, (2015): 62(6), 750–761.
44. Moradian, M. M., Beglaryan, D., Skozylas, J. M., & Kerikorian, V... Complete mitochondrial genome sequence of three tetrahymena species reveals mutation hot spots and accelerated nonsynonymous substitutions in ymf genes. PLOS ONE, 2,(2007).
45. Sueoka, N.. Translation-coupled violation of parity rule 2 in human genes is not the cause of heterogeneity of the dna g+c content of third codon position. *Gene (Amsterdam)*, (1999):238(1), 0–58.
46. Sueoka, N.. Directional mutation pressure and neutral molecular evolution. Proceedings of the National Academy of Sciences of the United States of America, (1988):85(8), 2653–2657.
47. Sharp, P.M., & Li, W. H.. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research, (1987):15(3), 1281–95.
48. Stamatakis, A.. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, (2014):30(9), 1312–1313.
49. Ronquist, F., Teslenko, M., Van, d. M. P., Ayres, D. L., Darling, A., & Hohna, S., et al.. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Systematic Biology, (2012):61(3), 539.

Tables

Table 1 The relative synonymous codon usage frequency (RSCU) of *S. lemnae* and other model organism

Animo acid	Codon	RSCU						Animo acid	Codon	RSCU					
		<i>S.lemnae</i>	<i>H.</i>	<i>C.</i>	<i>S.cerevisiae</i>	<i>P.</i>	<i>T.</i>			<i>S.lemnae</i>	<i>H.</i>	<i>C.</i>	<i>S.cerevisiae</i>	<i>P.</i>	<i>T.</i>
		<i>sapiens.</i>	<i>elegans</i>	<i>caudatum</i>		<i>thermophila</i>	<i>sapiens</i>			<i>elegans</i>	<i>caudatum</i>		<i>thermophila</i>		
Phe	UUU	<u>1.17</u>	0.94	0.72	<u>1.08</u>	<u>1.30</u>	<u>1.50</u>	Ser	UCU	<u>1.48</u>	<u>1.12</u>	<u>1.47</u>	<u>1.83</u>	<u>1.45</u>	<u>1.71</u>
	UUC	0.83	1.06	1.28	0.92	0.70	0.50		UCC	0.42	1.28	0.98	<u>1.09</u>	0.41	0.21
Leu	UUA	<u>1.55</u>	0.47	0.45	<u>1.60</u>	<u>2.18</u>	<u>2.34</u>		UCA	<u>1.76</u>	0.91	<u>1.44</u>	<u>1.16</u>	<u>2.07</u>	0.86
	UUG	<u>1.06</u>	0.78	<u>1.35</u>	<u>2.08</u>	0.91	<u>1.32</u>		UCG	0.22	0.33	0.83	0.51		0.21
	CUU	<u>1.39</u>	0.80	<u>1.86</u>	0.66	<u>1.82</u>	0.92	Pro	CCU	<u>1.58</u>	<u>1.14</u>	0.52	<u>1.18</u>		<u>1.67</u>
	CUC	0.57	1.15	1.38	0.28		0.41		CCC	0.43	1.29	0.23	0.54	2.00	0.33
	CUA	0.99	0.43	0.34	0.79	0.91	0.81		CCA	<u>1.82</u>	<u>1.10</u>	<u>2.75</u>	<u>1.88</u>	<u>2.00</u>	<u>2.00</u>
	CUG	0.44	2.36	0.63	0.59	0.18	0.20		CCG	0.17	0.47	0.51	0.39		
Ile	AUU	<u>1.4</u>	<u>1.10</u>	<u>1.52</u>	<u>1.47</u>	0.93	<u>1.35</u>	Thr	ACU	<u>1.95</u>	<u>1.01</u>	<u>1.34</u>	<u>1.50</u>	<u>1.54</u>	<u>2.00</u>
	AUC	0.63	1.38	1.23	0.89	0.53	0.31		ACC	0.54	1.39	1.02	0.97	0.62	0.38
	AUA	0.96	0.52	0.25	0.63	1.53	1.35		ACA	<u>1.35</u>	<u>1.15</u>	<u>1.15</u>	<u>1.06</u>	<u>1.54</u>	<u>1.63</u>
Met	AUG	1			1.00				ACG	0.16	0.46	0.49	0.47	0.31	
Val	GUU	<u>1.53</u>	0.89	<u>1.67</u>	<u>1.73</u>	<u>1.67</u>	<u>2.13</u>	Ala	GCU	<u>1.86</u>	<u>1.05</u>	<u>1.64</u>	<u>1.73</u>	<u>1.71</u>	<u>2.25</u>
	GUC	0.67	1.06	1.11	0.96	0.67	0.50		GCC	0.55	1.59	1.06	0.97	0.57	
	GUA	<u>1.19</u>	<u>1.60</u>	0.52	0.66	1.00	0.88		GCA	<u>1.43</u>	0.92	0.99	0.95	<u>1.14</u>	<u>1.50</u>
	GUG	0.62	0.44	0.70	0.65	0.67	0.50		GCG	0.16	0.44	0.31	0.35	0.57	0.25
Tyr	UAU	<u>1.31</u>	0.90	0.97	<u>1.02</u>	<u>1.33</u>	<u>1.61</u>	Cys	UGU	0.94	0.94	1.14	1.86	1.67	1.14
	UAC	0.69	1.10	1.03	0.98	0.67	0.39		UGC	<u>1.06</u>	<u>1.06</u>	0.86	0.14	0.33	0.86
Gln	UAA	<u>1.55</u>		<u>1.55</u>				Trp	UGG	1.00			1.00		
	UAG	0.87		0.87				Glu	GAA	<u>1.24</u>	0.86	<u>1.15</u>	<u>1.58</u>	<u>1.89</u>	<u>1.60</u>
Lys	AAA	<u>1.17</u>	0.88	0.84	1.05		<u>1.24</u>		GAG	0.76	1.14	0.85	0.18	0.11	0.40
	AAG	0.83	1.12	1.16	0.95	0.29	0.76	Arg	AGA	<u>4.4</u>	<u>1.29</u>	<u>1.79</u>	<u>4.24</u>	<u>3.00</u>	<u>3.75</u>
Asp	GAU	<u>1.51</u>		<u>1.36</u>	<u>1.46</u>	<u>1.64</u>	<u>1.56</u>		AGG	0.95	1.27	0.26	0.23	2.00	
	GAC	0.49		0.64	0.54	0.36	0.44		CGU	0.19	0.48	1.84	0.97	1.00	0.75
Gly	GGU	<u>1.15</u>	0.64		<u>2.84</u>	1.00	0.47		CGC	0.12	1.10	0.73	0.04		0.75
	GGC	0.69	1.35		0.24		0.47		CGA	0.29	0.65	1.07	0.42		0.75
	GGA	<u>1.76</u>	1.01		0.61	<u>3.00</u>	<u>3.06</u>		CGG	0.05	1.22	0.31	0.10		
	GGG	0.4	1.00		0.31			Asn	AAU	<u>1.42</u>	0.96	<u>1.10</u>	<u>1.11</u>	<u>1.77</u>	<u>1.43</u>
Gln	CAA	<u>1.08</u>	0.54		<u>1.46</u>		1.00		AAC	0.58	1.06	0.90	0.89	0.23	0.57
	CAG	0.5	1.46		0.54		1.00	Ser	AGU	<u>1.32</u>	0.91	0.76	0.95	<u>1.66</u>	<u>1.61</u>
His	CAU	<u>1.44</u>	0.85		<u>1.20</u>	<u>1.60</u>	<u>2.00</u>		AGC	0.8	1.44	0.52	0.30	0.41	1.39
	CAC	0.56	1.15		0.80	0.40									
Ter	UGA	1.00		0.59											
	UAA			1.56											
	UAG			0.60											

^a The 'RSCU' value represents the pattern of relative synonymous codon usage

Table 2 Preferred codons in genome of *S. lemnae*

Animo acid Codon RSCU ^a RSCU				Animo acid Codon RSCU RSCU				Animo acid Codon RSCU RSCU			
		(high)	(low)			(high)	(low)			(high)	(low)
Phe	UUU	<u>1.5</u>	0.66	Ser	UCU	<u>1.62</u>	0.74	Arg	AGA	<u>4.64</u>	0.92
	UUC	0.5	<u>1.34</u>		UCC	0.18	<u>1.29</u>		AGG	0.68	<u>1.18</u>
Leu	UUA	<u>2.41</u>	0.51	Pro	UCA	<u>2.09</u>	0.7	Gly	GGU	0.98	<u>1.19</u>
	UUG	0.96	<u>1.14</u>		UCG	0.04	0.7		GGC	0.36	<u>1.95</u>
	CUU	<u>1.5</u>	0.51		CCU	<u>1.28</u>	0.82		GGA	<u>2.52</u>	0.39
	CUC	0.19	0.93		CCC	0.27	0.62		GGG	0.15	0.47
	CUA	0.82	0.34		CCA	<u>2.41</u>	<u>1.13</u>	Lys	AAA	<u>1.37</u>	<u>1.3</u>
	CUG	0.12	<u>2.58</u>		CCG	0.04	<u>1.44</u>		AAG	0.63	0.7
Ile	AUU	<u>1.61</u>	0.66	Thr	ACU	<u>2.55</u>	0.49	Asp	GAU	<u>1.85</u>	0.79
	AUC	0.41	<u>1.99</u>		ACC	0.29	<u>1.6</u>		GAC	0.15	<u>1.21</u>
	AUA	0.99	0.35		ACA	<u>1.13</u>	<u>1.2</u>	Glu	GAA	<u>1.55</u>	<u>1.22</u>
	AUG	1	1		ACG	0.03	0.71		GAG	0.45	0.78
Met	AUG	1	1	Ala	GCU	<u>2.27</u>	0.9	Ser	AGU	<u>1.43</u>	0.64
	GUU	<u>2.24</u>	<u>1.03</u>		GCC	0.25	<u>1.32</u>		AGC	0.64	<u>1.93</u>
Val	GUC	0.31	0.98		GCA	<u>1.45</u>	<u>1.06</u>	Arg	CGU	0.18	<u>1.13</u>
	GUA	<u>1.01</u>	0.83		GCG	0.03	0.72		CGC	0	<u>1.36</u>
	GUG	0.44	<u>1.16</u>	Cys	UGU	0.91	0.58		CGA	0.45	0.37
	UAU	<u>1.68</u>	0.64		UGC	<u>1.09</u>	<u>1.42</u>		CGG	0.05	<u>1.05</u>
Tyr	UAC	0.32	<u>1.36</u>	TER	UGA	1	1	Asn	AAU	<u>1.71</u>	0.53
	UAA	<u>2.16</u>	0.32		UGG	1	1		AAC	0.29	<u>1.47</u>
Gln	UAG	0.64	0.37	Trp	CAA	<u>1.02</u>	<u>1.22</u>				
	UAA	<u>2.16</u>	0.32		CAG	0.18	<u>2.09</u>				
His	CAU	<u>1.76</u>	0.84	Gln							
	CAC	0.24	<u>1.16</u>								

^a The 'RSCU' value represents the pattern of relative synonymous codon usage

Table 3 Summary of correlation analysis of nucleotide composition and ENC

	A	C	T	G	GC	A3s	C3s	T3s	G3s	GC3s
%A	<u>1.000**</u>	<u>-.583**</u>	<u>-.408**</u>	<u>-.331**</u>	<u>-.619**</u>	<u>.640**</u>	<u>-.568**</u>	<u>-.156**</u>	<u>.166**</u>	<u>-.346**</u>
%C	<u>-.583**</u>	<u>1.000**</u>	<u>-.238**</u>	<u>.115**</u>	<u>.771**</u>	<u>-.544**</u>	<u>.820**</u>	<u>-.313**</u>	<u>-.054**</u>	<u>.616**</u>
%T	<u>-.408**</u>	<u>-.238**</u>	<u>1.000**</u>	<u>-.466**</u>	<u>-.465**</u>	<u>-.069**</u>	<u>-.186**</u>	<u>.688**</u>	<u>-.480**</u>	<u>-.448**</u>
%G	<u>-.331**</u>	<u>.115**</u>	<u>-.466**</u>	<u>1.000**</u>	<u>.722**</u>	<u>-.283**</u>	<u>.218**</u>	<u>-.348**</u>	<u>.451**</u>	<u>.455**</u>
%G+C	<u>-.619**</u>	<u>.771**</u>	<u>-.465**</u>	<u>.722**</u>	<u>1.000**</u>	<u>-.561**</u>	<u>.711**</u>	<u>-.441**</u>	<u>.252**</u>	<u>.721**</u>
ENC	.005	<u>.101**</u>	<u>-.105**</u>	<u>.024**</u>	<u>.086**</u>	<u>-.232**</u>	<u>.106**</u>	<u>-.210**</u>	<u>.374**</u>	<u>.318**</u>

** . Correlation is significant at the 0.01 level

*. Correlation is significant at the 0.05 level

Table 4 Summary of correlation between the first two axes and nucleotide constraints in

S.lemnae genome

Base composition	Axis1	Axis2
A	<u>.075</u> **	-.012
C	<u>-.061</u> **	<u>-.051</u> **
T	-.008	<u>.079</u> **
G	<u>-.036</u> **	<u>-.032</u> **
GC	<u>-.066</u> **	<u>-.056</u> **
A3s	<u>.051</u> **	<u>.028</u> **
C3s	<u>-.060</u> **	<u>-.048</u> **
T3s	.004	<u>.068</u> **
G3s	.014	<u>-.050</u> **
GC3s	-.039**	<u>-.069</u> **
ENC	-.045**	<u>-.065</u> **

** . Correlation is significant at the 0.01 level

*. Correlation is significant at the 0.05 level

Table 5 Correlation analysis among AROMO, GRAVY, the first two axes, GC 3s, ENC and GC in the polyprotein-coding region of *S. lemnae*

	Axis1	Axis2	ENC	GC3s	GC
Gravy r	<u>-.576</u> **	<u>.547</u> **	-.009	-.001	.010
P	.000	.000	.218	.850	.155
Aromo r	<u>-.409</u> **	<u>.484</u> **	.000	-.004	-.005
P	.000	.000	.968	.602	.452

** . Correlation is significant at the 0.01 level

*. Correlation is significant at the 0.05 level

Figures

SSU rDNA

ML/BI

0.1

Figure 1

To determine the phylogenetic relationship of *S. lemnae*, protozoan and metazoans, a phylogenetic tree was drawn.

Figure 2

A: A plot of PC1 and PC2 showed important features of the codon usage pattern in *S. lemnae* genes. B: In order to characterize the codon usage patterns from different types of genes, ribosomal related genes was statistical analysis by PCA.

Figure 3

The association analysis between ENC - GC3.

Figure 4

Neutrality plot analysis was used to estimated the relationship between GC12 and GC3, and it elucidates the extent of influence of mutation-selection equilibrium in shaping the CUB.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementarytableS2.doc](#)
- [Fig.S1.png](#)

- [Fig.S2.png](#)