A comparative computational study of the ‘rbcL’ gene in plants and in the three prokaryotic families—Archaea, cyanobacteria and proteobacteria

Sunil Kanti Mondal1,2*, Subhadeep Shit3# and Sudip Kundu1*

1Department of Biophysics, Molecular Biology and Bioinformatics, University of Calcutta, Kolkata 700 009, India
2Department of Biotechnology, The University of Burdwan, Rajbati, Burdwan 713 104, India
3School of Biotechnology and Life Sciences, Haldia Institute of Technology, Haldia 721 657, India

The rbcL (ribulose-1,5-biphosphate carboxylase oxygenase) gene plays a crucial role in carbon fixation. Previous studies shed light on its evolutionary relationship among different Phyla. Here, authors have done a comparative study of rbcL genes among proteobacteria, archaea, cyanobacteria and plants based on their compositional variations (GC%, amino acid frequency, codon usage, etc.). In addition we have checked the mutational pressure on rbcL genes. The results indicate that the rbcL genes of cyanobacteria have a wide range of GC%. On the other hand, those of the proteobacteria have mainly higher GC%. Preferences of some amino acids usages have observed in rbcL genes among all species with an exception of plant. Analysis of RSCU (relative synonymous codon usage) values depicts GC ending codon biasness in proteobacteria, archaea and with few exceptions in cyanobacterial species. On the other hand, AU ending codon biasness has been observed for plants. The correspondence analysis shows the significant difference in codon usage pattern among the selected four groups of species. The ENc (expected effective number of codons) plot implies the choice of rbcL gene codon is constrained only by mutational biasness. Moreover, the rbcL genes’ expression ability, as predicted by CAI (codon adaptation index), is similar for most of the species from different groups.

Keywords: Codon adaptation index (CAI), expected effective number of codons (ENc), ribulose-1,5-biphosphate carboxylase oxygenase, relative synonymous codon usage (RSCU)

Introduction

Ribulose-1,5-biphosphate carboxylase-oxygenase (E.C. 4.1.1.39, Rubisco), one of the most abundant enzymes on the globe, is responsible for catalyzing CO₂ assimilation to organic carbon via Calvin cycle. In its most prevalent conformation, found in most proteobacteria, cyanobacteria, algae and higher plants, Rubisco occurs as a hexadecamer composing of eight large subunit (rbcL, MW ≈ 56,000) and eight small subunits (rbcS, MW ≈ 14,000), assembling into an L₈S₈ holoenzyme. The large subunit plays the crucial role of carbon fixation. In Cyanobacteria and other prokaryotes, genes for both rbcL and rbcS are chromosomally encoded and co-transcribed. As evolutionary rate of rbcL is suitable for study of phylogeny, it is often used as model for phylogenetic investigation. Understanding of these evolution patterns may shed light on functional/structural features governing Rubisco activity.

Previous investigations on rbcL genes suggest a biphyletic origin of phototrophic eukaryotes. It is also found that rbcL of green algae/plant lineages is derived from cyanobacteria and forms one rbcL lineage; the second rbcL lineage consists of the non-green algae and is derived from proteobacteria. This scenario of evolution is also well supported by molecular phylogenies of other chloroplast-encoded genes like psbA, tufA, atpB, ClpC, etc. Phylogenies based on these genes suggest that there is a single cyanobacterial ancestor of plastids. These resulted in a number of hypotheses explaining the apparently contradictory results. For example, (a) a lateral gene transfer of rbcLS genes may have occurred from a proteobacteria into the ancestor that gave rise to non-green plants; (b) a lateral transfer of rbcLS operon may have occurred into cyanobacterial ancestor that gave rise to non-green plants; or (c) two rbcLS operons may have been present in cyanobacterial-like ancestor (that gave rise to plastids) and different copies were retained in green versus nongreen lineages.

While a large number of investigations have already done to understand the ancestry of Rubisco...
and its evolutionary relationship among different Phyla, present study deals with the \textit{rbcL} gene’s evolutionary pattern in prokaryotes (archaea, proteobacteria, cyanobacteria) and eukaryotes (plant); an extensive analysis on the compositional variation and similarity (GC content, amino acid usage frequency, codon bias, etc.) of the \textit{rbcL} genes; and also to understand the mutational pressure on \textit{rbcL} genes.

Materials and Methods

Collection of Data

Taxonomic and other related information, whole genome and 16S rRNA sequences of 43 species (10 Archaea spp., 10 Cyanobacteria spp., 15 Proteobacteria spp. & 8 Plant spp.) were collected from NCBI (http://www.ncbi.nlm.nih.gov). Annotated \textit{rbcL} gene and protein sequences were collected from the KEGG database (http://www.genome.jp/dbget-bin/).

Evolutionary Analysis

Bootstrapped trees of 16S rRNA and \textit{rbcL} gene sequences were generated using ClustalX (version 2) and PHYLIP (version 3.69)\textsuperscript{12,13}. Hierarchical clustering based on GC content was generated using DIANA\textsuperscript{14}.

Compositional Analysis

Parameters like GC content, RSCU (a measure of relative synonymous codon usage biasness) values, amino acid and tRNA frequencies were considered for compositional analysis. The GC, GC3s and Nc values\textsuperscript{15} were generated using CodonW (http://codonw.sourceforge.net/). These provided useful information regarding existence of mutational pressures acting on the genes\textsuperscript{16}. The expected effective number of codons (ENc) values from GC3s under H0 (null hypothesis, \textit{i.e.}, no selection) were calculated according to Equation 1, where S denotes GC3s.

\[
\text{ENc} = 2+S+ \left\{ 29/ \left[ S^2 + (1-S) ^2 \right] \right\} \quad \text{... (1)}
\]

Correspondence analysis was performed using CodonW to investigate major trend in RSCU variation among the genes. It identifies major trends in the variation of synonymous codon usage data and distributes genes along continuous axes in accordance with these trends. RSCU value close to 1 indicates lack of biasness, while much higher and lower values indicate preference and avoidance of that particular codon, respectively.

Types of motifs in gene sequence were calculated using the online tool MOTIFSCAN (http://hits.isb-sib.ch/cgi-bin/PFSCAN) and were aligned determining whether each of the genes are conserved among groups.

Statistical Significance Test

A statistical significance (\textit{z}) test was performed based on G and C ending codons of amino acids having minimum four degenerative codons with variation only in the third position to understand the preference of either codons on the respective organisms group.

\[
Z = (X_i - X_g)/\sqrt{(S_i^2+S_g^2)/(N_i+N_g)} \quad \text{... (2)}
\]

Here, \(X_i\) and \(S_i\) denote the average and standard deviation of \textit{i} (\textit{i}=C & G) ended codon’s RSCU values, respectively. \(N\) indicates the sample size.

Expressional Probability

Codon adaptation index (CAI), a measure of gene’s probable expression, was calculated\textsuperscript{17}. The codon preference model was also generated to check an existence of mutational biasness as well as also to understand the preferential effect over the different amino acids.

Results and Discussion

Phylogenetic Analysis

Phylogenetic trees generated based on \textit{rbcL} genes and 16S rRNA sequences are shown in Fig. 1. The main cluster of phylogenetic tree of the \textit{rbcL} genes of prokaryotes (Fig. 1a) has two sub clusters; one of archaeal and the other of \(\alpha\)- and \(\beta\)-proteobacterial species. It is joined by a cluster of \(\gamma\)-proteobacterial and finally preceded by cyanobacterial species. The phylogenetic tress of 16S rRNA sequences is presented in Fig. 1b. Its main cluster comprises of archaeal and cyanobacterial species as sub clusters, which is followed by cluster of \(\beta\)- and \(\gamma\)-proteobacterial, and of \(\alpha\)-proteobacterial species. Comparison of both phylogenetic trees shows that \textit{rbcL} genes of archaea and proteobacterial species have more similarity. On the other hand, 16S rRNA based tree clearly shows genetical similarity of archaeal and cyanobacterial species. The higher similarity among the \textit{rbcL} genes of cyanobacteria and plant species is evident from the phylogenetic tree of both the prokaryotic and plant species (Fig. 1c), and confirms about the evolutionary relationships as already observed by several previous studies.
Fig. 1 (a-c)—(a) Phylogenetic tree based on the \textit{rbcL} gene sequences; (b) Phylogenetic tree based on the 16SrRNA sequences of the species; & (c) Phylogenetic tree of both prokaryotes and eukaryotes based on the \textit{rbcL} gene sequences.
Compositional Variability

Variation in rbcL Gene

The hierarchical cluster based on GC content of the rbcL genes is presented in Fig 2. It has two main clusters. The first cluster forms two sub clusters: Cluster A and B, while the other cluster comprises of three sub clusters: Cluster C, D and E. Cluster A contains plant, whereas cluster B contains mainly cyanobacterial and few archaea species (Staphylothermus marinus, Methanosarcina mazei & Methanoculleus marisnigri). These two clusters have species possessing rbcL genes of low GC content. The cluster C represents species having comparatively higher GC content (56.42%) in their rbcL genes and it has species of different groups, such as, cyanobacteria (Trenocyechnecoccus elongatus & Synechococcus sp.), archaea (M. barkeri, Archaeoglobus fulgidus & Thermococcus gammatolerens) and proteobacteria (Acidithiococcus ferrooxidans & Nitrosomonas eutropha). However, the species having highest GC content in rbcL genes are present in cluster D and E, comprising mainly of the proteobacteria species and few cyanobacteria (Synechococcus sp. A-prime & Synechococcus sp. RCC307) and archaea (Thermofilum pendens & Natronomonas pharaonis). The results indicate that the rbcL genes of cyanobacterial species have a wide range of GC contents, whereas those of the proteobacterial species have mainly higher GC contents. It has been observed that the range of GC content in green plants varies from 28 to 42\%. It is also evident from the present results that the rbcL genes of plants have comparatively lower GC contents.

The gene has also been analyzed based on the physico-chemical properties of the amino acids present in its sequence. Amino acid frequencies are calculated and presented in three groups: amino acids with hydrophobic side chains in cluster I, polar amino acids with neutral side chains in cluster II and polar amino acids with charged groups in cluster III (Fig. 3). Amino acids like alanine, valine, isoleucine and leucine from cluster I show higher amino acid usage among all the four groups of species. Similarly, amino acids like glycine, proline and threonine from the cluster II show comparative higher usage ratio. On the other hand, glycine and cystiene usages are maximum and minimum, respectively among all the four groups of species. All the polar amino acids with charged groups, except histidine, (cluster III) show comparatively higher usage ratio in all groups of species. However, few exceptions are observed in a plant species, namely, Nicotiana benthamiana; it shows higher usage ratio of arginine (from cluster III), cystein and serine (from cluster II).

An intra and inter group correlation check was done based on the rbcL genes’ RSCU (Table 1) and the results show that proteobacterial and archaeal species have greater RSCU values in the GC-ending
codons, indicating greater usage of these codons. Cyanobacterial species also showed greater values for GC-ending codons, except amino acids arginine, alanine, valine, glutamic acid, glycine and lysine, while plant species showed greater RSCU values in the AU-ending codons, except amino acids asparagine and histidine.

**Codon Preference Check**

The rbcL genes’ RSCU values also lead to the establishment of a codon preference model, which thereby shows preference towards GC-ending codons by the prokaryotes, signifying a mutational bias towards G+C, but no such case is observed in the plant species (Table 1). A preference check of G3
over C3 or vice versa (Table 2) shows few significant results; archaeal species show biasness of C3 over G3 in all codons, except arginine and proline. Among cyanobacterial species, preference of G3 over C3 are found in amino acids like leucine, while C3 over G3 are seen in alanine, glycine, proline, threonine and serine. Among proteobacteria species, biasness towards C3 ending codons is observed, except amino acids leucine and proline. In plant species, preference exists in AU-ending codons, except asparagine and histidine showing a mild higher RSCU values in C3 ending codons.

**Statistical Significance Test**

Some amino acids like alanine and proline show no significant difference (P<0.05) between G and C ending codons both in archaea and plant species, while threonine shows significant difference in archaea, cyanobacteria, proteobacteria and plant species.

**Motif Analysis**

The general sequence of the rbcL motif is “G-x-[D]-F-x-K-x-D-E”. Degree of conservation using motif analysis shows that rbcL motifs of the different species are found almost conserved in all the species, except *N. eutropha*, *Brassica napus*, *N. benthamiana*.
and Gossypium raimondii (Fig. 4). Within the groups, cyanobacterial species show the most conserved sequence, except Prochlorococcus marinus where phenyl alanine residue has changed to leucine. This change has also been detected in archaea where in few species (Hyperthermus butylicus, S. marinus, T. gammatolerans & 3 Methanosarcina species), phenyl alanine residue has changed either to leucine or tyrosine. Proteobacteria species also show a conserved rbcL motif sequence.

### Expressional Variability

#### ENc Plot Analysis

The ENc plot analysis (ENC plotted against % (G+C)) was used to investigate patterns of synonymous codon usage, which shows that all the species lie below the expected curve, except two plant species B. napus and N. benthamiana, which are lying on the curve of the predicted values. This implies that the choice of the rbcL gene codon of the different species is constrained only by a mutational biasness.

#### Correspondence Analysis

To determine the codon usage of rbcL gene among the four groups, correspondence analysis on the genes RSCU values of the 43 organisms was carried out by a standard procedure. The distribution of the rbcL gene from the 43 species on the first two major axes of the correspondence analysis is shown in Fig. 6. Genes are recognized based on their groups. Proteobacteria and plant species are separated along the first major axis, while cyanobacteria and archaea species are separated along the second major axis, thereby depicting significant difference in their codon usage pattern. A closer look depicts that archaea species (N. pharaonis & Methanoculleus marisnigri) and cyanobacteria species (Cyanobacteria A & B-Prime, Synechoccus RCC307 & T. elongates) show similar RSCU pattern with proteobacteria species. Similarly, P. marinus (NATL1A, MIT9211 & AS9601)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>1.43</td>
<td>2.11</td>
<td>4.25</td>
<td>0.72</td>
</tr>
<tr>
<td>Glycine</td>
<td>2.13</td>
<td>2.33</td>
<td>7.94</td>
<td>-0.97</td>
</tr>
<tr>
<td>Leucine</td>
<td>0.54</td>
<td>-2.43</td>
<td>-3.97</td>
<td>-4.68</td>
</tr>
<tr>
<td>Proline</td>
<td>-1.5</td>
<td>2.56</td>
<td>-4.49</td>
<td>-0.08</td>
</tr>
<tr>
<td>Arginine</td>
<td>-5.42</td>
<td>-0.42</td>
<td>6.91</td>
<td>-2.61</td>
</tr>
<tr>
<td>Serine</td>
<td>4.11</td>
<td>3.68</td>
<td>-0.86</td>
<td>4.49</td>
</tr>
<tr>
<td>Threonine</td>
<td>2.43</td>
<td>4.89</td>
<td>7.12</td>
<td>2.98</td>
</tr>
<tr>
<td>Valine</td>
<td>2.3</td>
<td>-0.69</td>
<td>-0.38</td>
<td>-1.33</td>
</tr>
</tbody>
</table>

Amino acids having more than two codons have been considered. Z values greater than or less than 1.96 (at P<0.05) are significant.
species of cyanobacteria are found to have closer resemblance to all the plant species, except B. napus, N. benthamiana and G. raimondii, which are similar to the S. marinus and M. barkeri from archaea species.

**Codon Adaptation Index Variation**

Codon Adaptation Index (CAI) predicting the degree of expression of the *rbcL* gene analyzed shows an average rate of expression of 0.72-0.76 (Fig. 7). This indicates that *rbcL* is highly expressible gene and is playing an important role in these species life. Only three species from plant, *Helianthus annuus*, *Arabidopsis thaliana* and *Chlamydomonas reinhardtii* are found showing a comparative lower CAI values. This lower value might be due to usage of other Rubisco genes.

**Conclusion**

The comparative study of *rbcL* genes among different groups of species highlighted few significant outcomes. First of all, the results show that *rbcL* genes of cyanobacteria and plants are similar in comparison to the other groups, which is an already known evolutionary relationship. In addition to that, it further shows that *rbcL* genes of these two groups have lower GC% in comparison to those of proteobacteria and archaeal species, which show higher GC% values. On the other hand, cyanobacteria and the other prokaryotic groups show differences from plants on the basis of their higher usage ratio in GC preference codons, thereby signifying the existence of mutational pressure among their *rbcL* genes. Such mutational pressure is not observed in case of plants. Though the *rbcL* gene expression is nearly similar among the four groups but the pattern in their codon usages is found different.

**Acknowledgement**

Authors wish to thank the Department of Biotechnology, University of Burdwan, Burdwan for providing the facility of Computational Biology laboratory.
References


12. Felsenstein J, PHYLIP: Phylogeny interference package (version 3.69) (Department of Genome Sciences and Department of Biology, University of Washington, Washington, USA) 1989, 164-166.


