Peer Review Report

Review Report on Comparative Genomics of Thermosynechococcaceae and Thermostichaceae: Insight into Codon Usage Bias

Original Research, Acta Biochim. Pol.

EVALUATION

Q 1   Please summarize the main findings of the study.

The study examines the nucleotide composition of 15 genomes from four genera within two cyanobacteria families: Thermosynechococcaceae (three genera) and Thermostichaceae (one genus). It identifies differences in codon usage between the genera, influenced by GC content and amino acid composition. The analysis reveals distinct patterns between thermophilic (13 genomes) and non-thermophilic (2 genomes) cyanobacteria, with thermophiles showing higher GC content in codon bases, suggesting an adaptation to high-temperature environments.

Q 2   Please highlight the limitations and strengths.

The study's strength lies in its use of multiple indices, such as the Codon Adaptation Index, Codon Bias Index, and Relative Synonymous Codon Usage, to evaluate codon usage bias. This diverse methodology enhances the reliability of the results and provides a comprehensive understanding of codon usage patterns.

However, the study's limitation is the small dataset of 15 genomes, which may not fully represent the diversity of cyanobacterial species within the examined families. Additionally, the dataset is biased towards thermophilic species, with only one non-thermophilic species (Pseudocalidococcus azoricus) included, potentially limiting the generalizability of the findings to other species or genera.

Q 3   Please comment on the methods, results and data interpretation. If there are any objective errors, or if the conclusions are not supported, you should detail your concerns.

Methods: Using only 15 genomes may limit the study's ability to capture the full diversity of codon usage patterns across cyanobacteria. Expanding the dataset to include more genomes would enhance the generalizability of the findings.

Results: The results are well-presented.

Data Interpretation: The conclusions based on the 15 genomes may not broadly apply to all cyanobacteria. The link between codon usage and thermotolerance should be interpreted cautiously, as alternative factors could also explain the observed patterns.

Check List

Q 4   Please provide your detailed review report to the editor and authors (including any comments on the Q4 Check List)

I have several questions for the authors and hope they can address them.

1. Can the dataset of 15 genomes be expanded to include more genera and species? For instance, the current GTDB database (release 09-RS220) includes 31 representative genomes (species) from the

Thermosynechococcaceae family and 9 from the Thermostichaceae family. Could these genomes be incorporated to broaden the analysis?

2. The differences in nucleotide composition (e.g., GC content, codon usage) between just four genera are difficult to interpret in terms of significance. Focusing solely on the Thermosynechococcaceae and Thermostichaceae families may overlook important insights from other cyanobacterial families that might exhibit distinct codon usage patterns. Why did the authors not include data from a wider range of cyanobacteria to provide a broader, null-model comparison?

3. To provide deeper insights into codon usage across the analyzed genera, the authors could investigate whether the observed differences arise from orthologous genes or genus-specific genes.

4. In the Introduction, the authors mention that Pseudocalidococcus azoricus is a newly discovered species affiliated to the Thermosynechococcaceae family according to NCBI Taxonomy. Since it has not yet been classified in GTDB, would P. azoricus also be assigned to the Thermosynechococcaceae family based on its genomic sequence using GTDB-Tk?

5. The genome of P. azoricus is significantly larger (3.5-3.7 Mb) than the other studied genomes (2.4-3.0 Mb). The authors could include this information in the manuscript.

6. Do thermophilic cyanobacteria from other families of the Cyanophyceae class show similar characteristics (e.g., high GC content) as those observed in the two studied families?

7. Lines 114-116: What sequence identity threshold was used for the dereplication of genomes into species? How were the representative genomes selected?

8. Lines 377-380: The authors state that "codon pair pattern classified Pseudocalidococcus as the most divergent group (Figure 4c) from the other three thermophilic genera, which is entirely different from the phylogenomic relationship (Figure 4a)." However, the trees in Figures 4a and 4c seem quite similar to me. Could the authors clarify this point?

Q 5   **Is the English language of sufficient quality?**

Yes.

Q 6   **Is the quality of the figures and tables satisfactory?**

Yes.

Q 7   **Does the reference list cover the relevant literature adequately and in an unbiased manner?**

Yes.

Q 8   **Are the statistical methods valid and correctly applied? (e.g. sample size, choice of test)**

Not Applicable.

Q 9 Are the methods sufficiently documented to allow replication studies?

No.

Q 10 Are the data underlying the study available in either the article, supplement, or deposited in a repository? (Sequence/expression data, protein/molecule characterizations, annotations, and taxonomy data are required to be deposited in public repositories prior to publication)

Yes.

Q 11 Does the study adhere to ethical standards including ethics committee approval and consent procedure?

Not Applicable.

Q 12 Have standard biosecurity and institutional safety procedures been adhered to?

Not Applicable.

QUALITY ASSESSMENT

		Rating
Q 13	Originality	■ ■ □ □ □
Q 14	Rigor	■ ■ ■ □ □
Q 15	Significance to the field	■ ■ □ □ □
Q 16	Interest to general audience	■ □ □ □ □
Q 17	Quality of the writing	■ ■ □ □ □
Q 18	Overall quality of the study	■ ■ □ □ □