Fottea, Olomouc, 24(1): 23–26, 2024
DOI: 10.5507/fot.2023.008

The CIMS (Cyanobacterial ITS motif slicer) for molecular systematics

Nicolas A. Labrada[1*], Callahan A. McGovern[1], Aimee L. Thomas[1], Anne C. Hurley[1], Marie R. Mooney[1] & Dale A. Casamatta[1]

[1] *Department of Biology, College of Arts and Sciences, University of North Florida, 1 UNF Drive, Jacksonville, FL–32224, USA; [*]Corresponding author e–mail: nlab@fastmail.com*

Abstract: The 16S–23S rRNA Internal Transcribed Spacer (ITS) is a commonly employed taxonomic marker in cyanobacterial systematics. Due to numerous challenges in articulating phylogenetic relationships within this ubiquitous, ancient lineage, a polyphasic approach including 16S rRNA sequence data, ecology, morphology, and ITS secondary structure analysis has become the standard. In particular, the ITS motifs are being utilized in the erection of novel and cryptic taxa. However, this is challenging as researchers must manually mine and parse sequence data to visually find and identify ITS structures. This painstaking process deters researchers from using ITS motifs, may lead to inconsistencies, and is a rather dry, tedious enterprise. Thus, we present a simple, user–friendly web–based application for help in finding and preparing the most common cyanobacterial motifs (e.g., the Box–B, D1–D1´, tRNAs, etc.). After extensive testing, we note that the most common motifs are recovered at ca. 97%. These motifs can then be easily exported into Mfold or other similar folding packages. We hope that this will both provide a valuable tool for researchers but will also facilitate new discoveries and allow for greater consistency in publishing ITS comparisons. The tool can be accessed at www.phylo.dev.

Key words: Folding motifs, rRNA, secondary structures, taxonomy, 16S

Introduction

Cyanobacteria are an ancient lineage and describing their evolutionary relationships have always been a challenge (McGovern et al. in press). In order to avoid ambiguous classifications/phylogenies that arise primarily from phenotypic plasticity, cryptic diversity, and paucity of informative morphological traits common in cyanobacteria, researchers employ a multi–faceted approach to classification (e.g., Casamatta et al. 2005; Johansen & Casamatta 2005; Mai et al. 2018). This so–called polyphasic approach includes molecular, ecological, and morphological data, and has become crucial to the accurate identification and description of cyanobacterial relationships.

In addition to 16S rRNA sequence data, the RNA secondary structures of conserved regions within the 16S rRNA–23S rRNA internal transcribed spacer (ITS) region have also been found to be phylogenetically informative molecular characters (Iteman et al. 2000; Johansen & Casamatta 2005) (Fig. 1). The cyanobacterial 16S–23S ITS region contains up to two tRNA genes, as well as several conserved sequence motifs and regions of RNA secondary structure (Iteman et al. 2000). Among these conserved domains are the transcriptional anti–terminator sites Box–A, which is a short conserved sequence domain, and Box–B, which is a stem–loop structure of variable sequence (Condon et al. 1995). Other secondary structure motifs analyzed in the context of cyanobacterial systematics include the D1–D1', V2, and V3 helices (Johansen et al. 2011; Řeháková et al. 2007). Among different taxa, these regions may differ in their nucleotide sequence lengths, and their corresponding secondary structures (Fig. 1) may vary in the length of the basal stem, the relative size and position of unilateral and/or bilateral bulges in the helix, and the relative size of the terminal loop (Baldarelli et al. 2022).

While 16S rRNA sequence data is valuable for resolving higher–level taxonomic relationships among cyanobacteria (i.e., in determining orders and families), high values of 16S rRNA sequence similarity among closely related cyanobacterial strains has limited its usefulness as a molecular character for distinguishing species (Řeháková et al. 2007). Compared to the 16S rRNA gene, the 16S–23S ITS region exhibits a higher degree of variation among species, and is thus useful for determining relationships among closely related strains characterized by highly similar 16S rRNA gene sequences (Johansen et al. 2011). ITS sequence and secondary structure data have also been utilized in phylogenetics research concerning other organisms, including fungi (Koetschan et al. 2014) and green algae (Buchheim et

al. 2011; Mai & Coleman 1997). With increased employment of ITS motifs, a panoply of these secondary folding structures has emerged for use in species descriptions and identifications (Kabirnataj et al. 2020; Brown et al. 2021; McGovern in press).

While phylogenetically informative, ITS folding patterns are not without some challenges. First, there is the issue of homology: bacteria are renown for potentially possessing multiple operons (Iteman et al. 2000; Espejo & Plaza 2018), and reconstructing relationships relies on comparing homologous, not analogous, regions. Some cyanobacterial lineages possess multiple operons (Engene & Gerwick 2011). Second, not all of the motifs are necessarily present in all operons. For example, some transcripts contain two tRNA genes ($tRNA^{Ile}$ and $tRNA^{Ala}$), while others may contain only one or neither. Third, accurate annotation of a given ITS motif may be impeded by the presence of multiple sequence regions which resemble or are identical to the expected sequence of the motif's conserved flanking regions. For example, it is possible to obtain very similar Box–B motifs even though the leading sequences are very different from each other (Baldarelli et al. 2022). Fourth, when assembling large ITS plates, it can be both tedious and error–prone to use Microsoft Word (commonly employed to manually highlight/annotate regions) in these endeavors.

To address these issues, we present CIMS (the Cyanobacterial ITS Motif Slicer). This web–based application has been created to find the most commonly used ITS folding motifs (e.g., D1–D1′, V2, etc.), both tRNAs, and other pertinent nucleotide regions (e.g., the leader, D2–D3, etc.) (Fig. 1). The CIMS script is also available for users to download and run on MacOS/Linux and Windows systems. To help researchers ensure they are using homologous operons (e.g., containing the same number of tRNAs) when comparing ITS secondary structures between taxa, the script can be executed with a flag (–t) that counts tRNAs in each ITS region and returns taxa and their corresponding tRNA count. Researchers may also choose to output only certain motifs (e.g., D1–D1′, Box–B, etc.) by using the flag (–s). Both the web–based CIMS application and the script accept FASTA files and GenBank accession numbers as input. CIMS then returns the sequences and lengths of all identified ITS motifs.

The program leverages the Biopython library (Cock et al. 2009) both to query the Entrez database and to parse the FASTA file. CIMS searches for the nucleotide sequence which marks the end of the 16S rRNA gene (CCTCCTT or CCTCCTA) in order to identify where the ITS region begins; the ITS region is then sliced off from the full sequence and the program begins looking for the conserved flanking regions of each element in consecutive order (Fig. 1).

CIMS will return the following features (if present in the transcript) in order: ITS leader, D1–D1′, spacer D2–D3 spacer, $tRNA^{Ile}$, V2, $tRNA^{Ala}$, Box–B, Box–A, D4, and V3. With every successful find it slices off the substring of the motif and stores it in a dictionary where the key is the motif name, and the value is an array of all the possible sequences for that motif. Once the program is finished, it either downloads the output in the form of a text file to the user's computer (web–based) or prints the output to the terminal and saves the output as a JSON file in the working directory (if running the script directly). The program follows Entrez guidelines that require an email address when running a GenBank query, but the script **does not** store or transmit email addresses in any other way.

Results and Discussion

To assay the efficacy of the software, we performed extensive testing from all cyanobacterial orders except the Gloeobacteriales (due to a lack of appropriate, full ITS data available) employing ca. 300 taxa (Table 1). For full transcripts, CIMS was 100% effective at finding the tRNAs (or identifying that no tRNAs were present). Both D1–D1′ and Box–B motifs (the most commonly folded motifs in manuscripts) were recovered at >95%, but with some heterogeneity based on the order (Table 1). The differences between the orders are likely less a result of the software but rather based on sampling effort (some orders had much more sequence data available) (Table S1). A preliminary review of publications containing predicted V3 secondary structures for cyanobacterial taxa revealed that, in contrast to the D1–D1′ and Box–B regions, the V3 region had highly variable 5′ and 3′ flanking sequences. Therefore, given that CIMS currently identifies semi–conserved regions of secondary structure (e.g., Box–B and D1–D1′) by searching for specified sets of flanking sequences, CIMS's success rate in accurately identifying the V3 region has not yet been quantified and thus is not here presented.

Table 1. The efficacy of CIMS in returning the most commonly employed motifs by cyanobacterial order based on ca. 300 test cases of transcripts that represent full ITS regions. tRNAs were always recovered from ITS regions that contained them. Actual strains employed are in Suppl. Table 1.

Order	D1-D1′	BoxB	tRNAs
Synechococcales	96.49%	92.98%	100%
Chroococcales	100%	97.87%	100%
Nostocales	93.90%	92.68%	100%
Pleurocapsales	100%	100%	100%
Chroococcidiopsidales	93.75%	87.50%	100%
Spirulinales	94.44%	94.44%	100%
Oscillatoriales	100%	98.39%	100%
Cyanobacteria (all orders)	96.94%	94.84%	100%

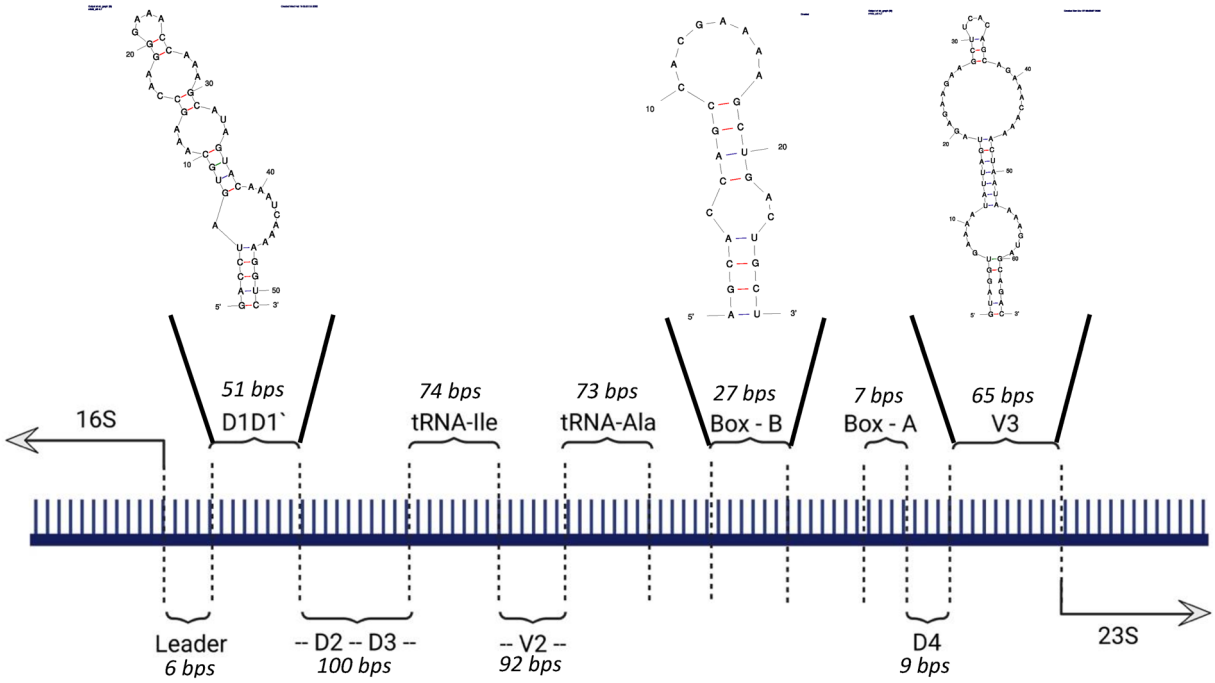

Fig. 1. ITS region of *Anagnostidinema visii* LHM–M between the 16S gene and the 23S gene and its motifs labeled in order. Secondary structures of motifs D1–D1`, Box–B, and V3 shown.

Potential limitations/confounding factors

While we are confident that CIMS will facilitate phylogenetic assessments, we realize that there exist some limitations with any script.

1. CIMS only looks for patterns that are specified in the script (e.g., looking for the flanking regions before and after the Box–B or D1–D1′, etc.); it is deterministic for any given sequence.
2. Right now there is no user–friendly configuration of the search parameters, e.g., if you want to modify the flanking regions (perhaps you know that your particular strain has a single nucleotide polymorphism), you have to go to the code itself to initiate this modified search.
3. GenBank is renowned for the lack of oversight concerning sequence quality, and CIMS cannot evaluate the quality of the sequence (e.g., is the sequence data accurate).
4. The CIMS script has been optimized for cyanobacteria; we cannot attest to the efficacy with other Bacteria.
5. There is no way of verifying if the motif or folded secondary structure is "correct" (no quality score). We suggest that users compare their structures with those from the literature.
6. CIMS may output several motif sequences when there are multiple matching flanking regions found. We suggest comparing the lengths and predicted secondary structures of these sequences to those of other cyanobacterial ITS regions presented in the literature; if one is similar in length and structure to others in the literature while another is radically different, the principles of parsimony may guide our explorations.
7. Primer biases ~~may~~ present some operons more commonly than others (e.g., Balderelli et al. 2022). This issue is broader than merely being a feature of CIMS but included here for completeness.

With those caveats, the useful aspects of this package are that it: 1) makes it easier to potentially find the motifs, 2) allows identification of operonic variants (i.e., how many tRNA's are present to look at homologous regions), 3) allows researchers easy access to potentially phylogenetically informative data (e.g., leader length, total bp, etc., all of which are being increasingly employed in manuscripts), and 4) enables consistent and automated annotation.

Future directions

This project is open source so that the cyanobacterial systematics community can benefit from it and, hopefully, contribute to the future growth of this application. As more researchers employ it, we will continue to modify the code for improved functionality.

As a tool for scientists, we always welcome feedback on positive and negative aspects of the script. On the GitHub website there is an "Issues" tab to provide feedback. We are also planning on several possible future features. First, HTML output of the ITS region color coded by motif (which is how most researchers visualize the sequence). Second, user specified flanking regions for motifs may be useful for lineages whose evolutionary trajectories have created novel or unique sequence variants. Third, we hope to be able to connect

the results to a folding software so that users can find and fold motifs in a more streamlined fashion.

Future work on the CIMS tool will also include improving its capability to identify the V3 region. The 5′ and 3′ flanking sequences of this region, which form the basal portion of the V3 stem–loop structure, were found to be more variable than those of the D1–D1′ and Box–B regions. Due to this variability, the CIMS script is not currently optimized to identify the V3 region across a wide range of taxa. In contrast, regions D1–D1′ and Box–B are characterized by more conserved 5′ and 3′ flanking sequences, and thus these regions were correctly identified in 96.94% and 94.84%, respectively, of all ITS sequences tested. While further work is required in order for CIMS to successfully identify V3 across a wide range of cyanobacterial taxa, the current CIMS script can be edited by users to search for any specified set of V3 flanking sequences. This may be useful in situations where the user has manually identified the V3 region in several ITS sequences of interest, and expects that the flanking sequences will be highly similar among all remaining ITS sequences of interest (for instance, when the user is annotating ITS sequences obtained from a group of very closely related strains or from multiple clones derived from a single strain).

The web–based tool can be accessed at www.phylo.dev. The CIMS script and documentation are available at: https://github.com/nlabrad/CIMS–Cyanobacterial–ITS–motif–slicer.

References

Baldarelli, L.M; Pietrasiak, N.; Osorio–Santos, K. & Johansen, J.R. (2022): *Mojavia aguilerae* and *M. dolomitestris* – two new Nostocaceae (Cyanobacteria) species from the Americas. – Journal of Phycology 58: 502–516.

Buchheim, M.A.; Keller, A.; Koetschan, C.; Förster, F.; Merget, B. & Wolf, M. (2011): Internal transcribed spacer 2 (nu ITS2 rRNA) sequence–structure phylogenetics: towards an automated reconstruction of the green algal tree of life. – PLoS one 6: e16931. DOI: 10.1371/journal.pone.0016931

Casamatta, D.A.; Johansen, J.R.; Vis, M.L. & Broadwater, S. (2005): Molecular and morphological characterization of ten polar and near-polar strains within the Oscillatoriales (Cyanobacteria). – Journal of Phycology 41: 421–438.

Cock, P.J.; Antao, T.; Chang, J.T; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B. & de Hoon, M.J. (2009): Biopython: freely available Python tools for computational molecular biology and bioinformatics. – Bioinformatics 25: 1422–1423.

Condon, C.; Squires, C. & Squires, C.L. (1995). Control of rRNA transcription in Escherichia coli. – Microbiological reviews 59: 623–645.

Espejo, R.T. & Plaza, N. (2018): Multiple ribosomal RNA operons in Bacteria: their concerted evolution and potential consequences on the rate of evolution of their 16S rRNA. – Frontiers in Microbiology 9: 1232.

Iteman, I.; Rippka, R.; Tandeau de Marsac, N. & Herdman, M. (2000): Comparison of conserved structural and regulatory domains within divergent 16S rRNA–23S rRNA spacer sequences of cyanobacteria. – Microbiology 146: 1275–1286.

Johansen, J.R. & Casamatta, D.A. (2005): Recognizing diversity through adoption of a new species paradigm. – Algological Studies 117: 71–93.

Johansen, J.R.; Kovacik, L.; Casamatta, D.A.; Fučiková, K. & Kaštovský, J. (2011). Utility of 16S–23S ITS sequence and secondary structure for recognition of intrageneric and intergeneric limits within cyanobacterial taxa: *Leptolyngbya corticola* sp. nov. (Pseudanabaenaceae, Cyanobacteria). – Nova Hedwigia, 92: 283–302.

Kabirnataj, S.; Nematzadeh, G.A.; Talebi, A.F.; Saraf, A.; Suradkar, A.; Tabatabaei, M. & Singh, P. (2020): Description of novel species of *Aliinostoc*, *Desikacharya* and *Desmonostoc* using a polyphasic approach. – International Journal of Systematic and Evolutionary Microbiology 70: 1–9.

Koetschan, C.; Kittelmann, S.; Lu, J.; Al–Halbouni, D.; Jarvis, G.N.; Müller, T.; Wolf, M. & Janssen, P.H. (2014): Internal transcribed spacer 1 secondary structure analysis reveals a common core throughout the anaerobic fungi (Neocallimastigomycota). – PloS one 9: e91928. DOI: https://doi.org/10.1371/journal.pone.0091928

Mai, J.C. & Coleman, A.W. (1997) : The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants. – Journal of Molecular Evolution 44: 258–271. DOI: https://doi.org/10.1007/pl00006143

Mai, T.; Johansen, J.; Pietrasiak, N.; Bohunická, M. & Martin, M.P. (2018): Revision of the *Synechococc ales* (Cyanobacteria) through recognition of four families including Oculatellaceae *fam. nov*. and Trichocoleaceae *fam. nov*. and six new genera containing 14 species. – Phytotaxa 365: 1–59.

Řeháková, K.; Johansen, J.R.; Casamatta, D.A.; Xuesong, L. & Vincent, J. (2007): Morphological and molecular characterization of selected desert soil cyanobacteria: three species new to science including *Mojavia pulchra* gen. et sp. nov. – Phycologia 46: 481–502.

Supplementary material

The following supplementary material is available for this article:

Table S1. All taxa used as test cases for CIMS (Box-B and D1-D1').

Table S2. ITS region and flanking sequences used in CIMS.

This material is available as part of the online article (http://fottea.czechphycology.cz/contents)

Received October 27, 2022
Accepted February 2, 2023