Repeat-type distribution in trnL intron does not correspond with species phylogeny: comparison of the genetic markers 16S rRNA and trnL intron in heterocystous cyanobacteria

Ilona Oksanen,^{1,2,3} Katileena Lohtander,³ Kaarina Sivonen¹
and Jouko Rikkinen^{2,3}

Correspondence

Ilona Oksanen
ilona.oksanen@helsinki.fi

Department of Applied Chemistry and Microbiology¹, Department of Ecology and Systematics²
and Department of Applied Biology³, University of Helsinki, 00014 University of Helsinki,
Finland

tRNA^{Leu} UAA (trnL) intron sequences are used as genetic markers for differentiating cyanobacteria and for constructing phylogenies, since the introns are thought to be more variable among close relatives than is the 16S rRNA gene, the conventional phylogenetic marker. The evolution of trnL intron sequences and their utility as a phylogenetic marker were analysed among heterocystous cyanobacteria with maximum-parsimony, maximum-likelihood and Bayesian inference by comparing their evolutionary information to that of the 16S rRNA gene. Trees inferred from the 16S rRNA gene and the distribution of two repeat classes in the P6b stem-loop of the trnL intron were in clear conflict. The results show that, while similar heptanucleotide repeat classes I and II in the P6b stem-loop of the trnL intron could be found among distant relatives, some close relatives harboured different repeat classes with a high sequence difference. Moreover, heptanucleotide repeat class II and other sequences from the P6b stem-loop of the trnL intron interrupted several other intergenic regions in the genomes of heterocystous cyanobacteria. Cluster analyses based on conserved intron sequences without loops P6b, P9 and parts of P5 corresponded in most clades to the 16S rRNA gene phylogeny, although the relationships were not resolved well, according to low bootstrap support. Thus, the hypervariable loop sequences of the trnL intron, especially the P6b stem-loop, cannot be used for phylogenetic analysis and conclusions cannot be drawn about species relationships on the basis of these elements. Evolutionary scenarios are discussed considering the origin of the repeats.

INTRODUCTION

The UAA anticodon of the tRNA^{Leu} gene is, in many cyanobacteria, interrupted by two copy types of group I introns, type I containing the introns of chloroplasts (e.g. Kuhsel *et al.*, 1990; Xu *et al.*, 1990; Paquin *et al.*, 1997; Strehl *et al.*, 1999; Besendahl *et al.*, 2000) and type II the eubacterial introns (Rudi & Jakobsen, 2002; Veprikskiy

et al., 2002). The type I tRNA^{Leu} UAA intron (trnL intron), which is present only as one sequence type in each cyanobacterial strain (Paulsrud & Lindblad, 1998), has been used in taxonomic and ecological studies of cyanobacteria. Within heterocystous cyanobacteria, the host specificity of *Nostoc* strains (Paulsrud & Lindblad, 1998; Paulsrud *et al.*, 2000, 2001; Costa *et al.*, 1999, 2001; Oksanen *et al.*, 2002; Rikkinen *et al.*, 2002; Summerfield *et al.*, 2002; Linke *et al.*, 2003; Wirtz *et al.*, 2003) and the form species concept of *Nostoc commune* (Wright *et al.*, 2001) have been studied by comparing nucleotide sequences of the introns and by deriving taxonomic relationships of the similarities. trnL introns are highly variable in their seven loops and are conserved in the stem region. From *Nostoc* strains, two sequence classes of mostly double-stranded heptamers, 5'-TDNGATT-3' (class I) and 5'-NNTGAGT-3' (class II), have been described in different copy-numbers from the P6b stem-loop of the trnL intron (Costa *et al.*, 2002), causing nucleotide and size variation that is considered useful in strain identification. In previous phylogenetic

Published online ahead of print on 14 November 2003 as DOI 10.1099/ijs.0.02928-0.

Abbreviations: ML, maximum-likelihood; MP, maximum-parsimony; trnL, tRNA^{Leu} UAA.

Database accession numbers for the new 16S rRNA gene sequences are AY112693, AY112694, AY328896–AY328900 (GenBank) and for the trnL intron AY112695 (GenBank), AJ571702–AJ571720 (EMBL).

Bayesian tree of the 16S and trnL intron sequences, the origin of the strains studied and the list of the strains with no intron PCR products are available as supplementary material in IJSEM Online.

analyses, the most variable stem-loops, P6b and P9, have been included (Linke *et al.*, 2003; Wirtz *et al.*, 2003) or excluded either partially (Rudi & Jakobsen, 1999) or totally (Paquin *et al.*, 1997; Paulsruud & Lindblad, 1998; Besendahl *et al.*, 2000; Wright *et al.*, 2001) to improve the reliability of the sequence alignment. However, whether the distribution of the two repeat classes follows strain phylogeny has not been tested, nor whether trees based solely on the conserved regions of trnL intron sequences are congruent with those generated from conventional phylogenetic markers, such as the 16S rRNA gene (e.g. Turner *et al.*, 1999).

We constructed the 16S rRNA gene phylogeny of a variety of heterocystous cyanobacteria, including a range of symbiotic *Nostoc* strains, using maximum-parsimony (MP), maximum-likelihood (ML) and Bayesian inferences. To reveal whether strains with divergent trnL intron P6b stem-loop sequences are more distantly related to each other than strains with similar P6b sequences, we compared the 16S rRNA gene phylogeny with the distribution of repeat classes in the trnL intron sequences. We also evaluated how the results of cluster analyses based on the conserved intron sequences (without loops P6b, P9 and parts of P5) corresponded to the 16S rRNA gene phylogeny that was considered to reflect the species phylogeny of the cyanobacteria studied (e.g. Turner *et al.*, 1999). To reveal the frequency of repeats of the trnL intron, we searched for them on sequence databases.

METHODS

Cyanobacterial strains. Representatives of the heterocystous genera *Anabaena*, *Aphanizomenon* (Gugger *et al.*, 2002), *Calothrix* (Turner *et al.*, 1999; Lyra *et al.*, 2001), *Cylindrospermum*, *Nodularia* (Lehtimäki *et al.*, 2000; Lyra *et al.*, 2001) and *Nostoc* (Miao *et al.*, 1997; Lyra *et al.*, 2001; Rikkinen *et al.*, 2002) from cyanobacterial subsection IV (Castenholz, 2001) and *Tolypothrix* from subsection IV.II were included in final analyses (in Supplementary Table A; additional strains studied in Supplementary Table B). Sequences from non-heterocystous *Pseudanabaena* (Ishida *et al.*, 2001) and *Planktothrix* (Lyra *et al.*, 2001) from subsection III and *Synechococcus* (Ishida *et al.*, 2001) from subsection I were used as outgroups. Most of the *Nostoc* strains originate from lichen symbiosis. There is only one *Nostoc* strain from each lichen thallus, except for a *Lobaria pulmonaria*, from the thallus fragments of which two strains were cultured. Jiri Komarek kindly determined *Tolypothrix* cf. and *Nostoc* cf. *trichormus* strains based on morphology. Most of the sequences reported here originate from unialgal and axenic cultures. Axenicity was controlled using R2A medium (International Diagnostics Group).

DNA extraction and PCR. Total DNA was either extracted by the DNeasy tissue kit (Qiagen) with previous modifications (Lohtander *et al.*, 2000) or obtained directly from crushed cells. DNA polymerases used were *Taq* (Sigma), Platinum *Taq* (GibcoBRL) or Fast Start *Taq* (Roche). The trnL intron sequence was amplified in a nested PCR with two sets of primers: cyanobacterial specific primers A and C (Paulsruud & Lindblad, 1998) and TL25 and TL23 (Biniszkiewicz *et al.*, 1994). First amplification with primers A and C was carried out in 20 µl as described in Paulsruud & Lindblad (1998) with the addition of 1 M betaine (Sigma). DNA was re-amplified with primers TL25 and TL23 in 50 µl with 0.5 or 1.0 µl PCR

product and 1 M betaine at an annealing temperature of 60 °C for 1 min in 35 cycles following the instructions from the DNA polymerase manufacturer. Cyanobacterial 16S rRNA was amplified with primers PCR 18 and PCR1 (Wilmotte *et al.*, 1993) in 25 or 50 µl at an annealing temperature of 56 °C for 1 min for 30 cycles following the instructions from the DNA polymerase manufacturer.

DNA sequencing. PCR products were purified with Microcon-PCR filters (Millipore). The PCR product of the trnL intron was sequenced bidirectionally with primers TL25 and TL23 (Biniszkiewicz *et al.*, 1994). The 16S rRNA gene was sequenced with primers from Wilmotte *et al.* (1993). Cycle sequencing of the trnL intron and the 16S rRNA gene was performed using the ABI Prism dye terminator cycle sequencing ready reaction kit (Perkin-Elmer) and PCR products were purified on MicroSpin G-50 columns (Amersham Pharmacia Biotech). Sequence reactions were analysed with an ABI Prism 377 automated sequencer or with an ABI Prism 310 Genetic Analyzer (PE Biosystems).

Sequence alignment. The 16S rRNA gene and trnL intron sequences were aligned using the CLUSTAL W program (Thompson *et al.*, 1994). Alignments were refined manually and pairwise sequence identities were obtained with BioEdit 5.0.0 (available at <http://www.mbio.ncsu.edu/BioEdit/>). The GenBank database was queried with parts of the trnL intron P6b sequence using BLAST (Altschul *et al.*, 1997) and Cyanobase, with their BLAST search

Fig. 1. Hypothetical secondary structure of the trnL intron from *Nostoc* strain I.3. The loop sequence of P6b and of P9 and up to 15 nucleotides of P5 were excluded from the phylogenetic analyses, which included only the conserved intron stem and loops without length variation.

(<http://www.kazusa.or.jp/cyano/blast.html>). A secondary structure of a trnL intron from a *Nostoc* (Fig. 1) is based on the model of Cech *et al.* (1994).

Phylogenetic analyses. The trnL intron and 16S rRNA gene data were analysed separately and combined to reveal the phylogenetic utility of the trnL intron in resolving relationships among heterocystous cyanobacteria. Two sets of trnL intron sequences ($n=22$, $n=18$ strains) and one of 16S rRNA gene sequences ($n=41$ strains) were analysed (Table 1). The dataset with 41 strains was used in all other analyses except hypothesis testing. In the first trnL intron dataset, sequences of *Pseudanabaena* and *Synechococcus* were excluded from the MP, ML and Bayesian analyses, since their inclusion in these analyses resulted in unresolved trees. From the second trnL intron dataset, *Calothrix* and *Planktothrix* strains, these being most distantly related to *Nostoc*, were removed to test whether this would improve the resolution of the tree. In all intron alignments, the loop regions P6b, P9 and 15 nucleotides from the P5 loop (Supplementary Figure) were removed. In addition, only one representative was included from strains that shared identical intron sequence in phylogenetic analyses of the intron (Table 1; datasets 1 and 2). To ensure stability of resulting groupings, three different phylogenetic approaches with varying algorithms and a test of alternative hypotheses were applied to the sequence datasets. The support values, bootstrap proportion and the Bayesian posterior probabilities applied for phylogenetic groups measure different

parameters from the models. Bayesian posterior probabilities indicate the probability of the model (phylogenetic tree) on estimated priors, whereas the bootstrap analysis indicates the relative probability of the simulated data (sequences) on a given model (Alfaro *et al.*, 2003). Thus, the two support values do not necessarily always correlate (Suzuki *et al.*, 2002; Douady *et al.*, 2003). Substitution models can create additional biases in the analyses as different models are used: unweighted parsimony versus the best-fitting, weighted substitution models in the likelihood-based ML and Bayesian inference (Posada & Crandall, 2001).

Model testing. The best-fitting nucleotide substitution model was determined by testing 56 different models (modelblock3; Posada & Crandall, 2001) for each dataset in PAUP 4.0b10 (Swofford, 2002). The model was chosen with the MODELTEST 3.06 program using the Akaike Information Criterion (Posada & Crandall, 1998; Table 1).

MP analysis. MP analyses were conducted in equally weighted heuristic searches in PAUP 4.0b10a using the TBR branch-swapping algorithm and a random addition sequence with 1000 replicates (Swofford, 2002). Strict consensus trees were constructed from the most parsimonious trees. Non-parametric MP bootstrap analysis in PAUP incorporated random addition of sequences with replicates, and TBR branch-swapping. Random addition sequences of 100 replicates were used for the 16S rRNA, and combined sequence data and 1000 replicates for the trnL intron data.

Table 1. Parameters of the four datasets and results of the MP, ML and Bayesian phylogenetic analyses

Aligned nucleotide positions 64–1440 of the 16S rRNA gene (in *Nostoc* PCC 7120; NC003272) corresponded to positions 82–1494 in *Escherichia coli* (AE005174). The loop regions P6b and P9 and 15 nucleotides from P5 were excluded in trnL intron datasets 1 and 2 and in the combined dataset.

Analyses/datasets	16S rRNA gene	trnL intron core dataset 1	trnL intron core dataset 2	16S rRNA and intron core
Matrix length/number of strains	1385/41	202/22	201/18	1602/41
Substitution model*	GTR + G + I	TrN + I + G	TrN + I + G	GTR + G + I
Variable sites/PI†	358/259	56/39	42/34	439/322
No. of MP trees/length	2/937	21/126	6/81	8/1140
No. of ML trees/−lnL	1/6774·71685	1/883·2378	3/668·0133	1/8007·7052
A–C, A–G, A–T, C–G, C–T‡	1·08, 1·82, 1·55, 0·53, 2·62	1·00, 2·65, 1·00, 1·00, 7·28	1·00, 3·82, 1·00, 1·00, 10·65	1·02, 1·99, 1·58, 0·55, 3·14
Shape parameter	0·63	0·78	0·97	0·61
Proposed invariable sites	0·60	0·56	0·66	0·57
Frequency of A, C, G, T	0·25, 0·23, 0·32, 0·20	0·35, 0·17, 0·26, 0·21	0·35, 0·19, 0·27, 0·20	0·27, 0·22, 0·31, 0·20
ML constrained trees§	6668·6208			
Non-constrained trees§	6590·1152			
Bayesian trees with posterior probability of ≥99%	3642	8733	8438	2263
Bayesian trees with posterior probability of ≥95%	3442	8373	8078	2063
Shape parameter (mean)	0·37	0·06	0·06	0·46
Proposed invariable sites	0·52	0·39	0·53	0·52

*Best-fitting nucleotide substitution model chosen with AIC by MODELTEST 3.06 (Posada & Crandall, 2001).

†Number of parsimony-informative sites.

‡Constrained tree forced monophyly of the terrestrial *Nostocs* with repeat class I in trnL intron. Both analyses included 40 taxa.

§Substitution rate G–T is 1·0.

||Excluding the burn-in set (5000 generations for 16S rRNA and combined datasets, 1000 for intron data).

ML analysis. ML analyses were conducted in PAUP 4.0b10 (Swofford, 2002) using best-fitting substitution models with their associated parameters (Table 1) in heuristic searches with options TBR and random addition sequence with 30 replicates. ML non-parametric bootstrap analysis in PAUP incorporated TBR branch-swapping, starting trees by neighbour-joining analysis and 10 000 bootstrap replicates.

Bayesian inference. Bayesian inference was applied with MrBayes 3.0b3 and 3.0b4 (Huelsenbeck & Ronquist, 2001; Huelsenbeck *et al.*, 2001). MrBayes implements a Markov chain Monte Carlo method that runs many chains simultaneously. With the best-fitting nucleotide substitution model, two or three separate runs incorporating four chains were performed for a million generations, of which every 100th was sampled, generating a total of 10 000 trees each run. Runs started from a random tree. To ensure that the data were from stationary phase, likelihood scores were plotted on a diagram and unstable initial cycles were removed. The proportion of discarded initial cycles was 50 % in analyses of 16S rRNA gene sequences and in combined datasets of 16S rRNA gene sequences and trnL intron sequences, and 10 % in analyses of the trnL intron sequences. Majority rule (50 %) consensus trees were viewed with TreeView 1.6.6 (Page, 1996). The confidence of groupings, such as the paraphyly of *Nostoc* strains with the trnL intron repeat class I in the 16S rRNA gene tree, was studied from the taxon partition, which gave the number of trees in which each grouping showed and the posterior probabilities for these trees.

Shimodaira–Hasegawa test. An alternative hypothesis for the constructed 16S rRNA gene tree was tested with the one-tailed Shimodaira–Hasegawa log-likelihood test (Shimodaira & Hasegawa, 1999; Goldman *et al.*, 2000). The null hypothesis was a monophyletic origin of *Nostoc* strains with the class I repeat in the P6b stem-loop of the trnL intron. Thus, a constrained tree from the 16S rRNA gene dataset, in which all the *Nostoc* strains with class I repeats were forced to be monophyletic, was constructed with TreeView 1.6.6 (Page, 1996). Constrained and non-constrained sequence data ($n=40$, excluding strain *Calothrix* PCC 7102) were analysed with ML methods. The statistical difference between the constrained and non-constrained trees ($-\ln L_0=6668.6208$, $-\ln L_1=6590.1152$) was tested as implemented in PAUP with re-sampling estimated log-likelihood option (RELL) and 1000 bootstrap replicates.

RESULTS AND DISCUSSION

Evolution of trnL intron P6b stem-loop is incongruent with the species phylogeny

Conserved elements of the trnL intron have been used for the identification of cyanobacterial strains to the genus and species levels and the more variable elements for distinguishing between different *Nostoc* strains (Wright *et al.*, 2001; Paulsrud, 2001; Rikkinen *et al.*, 2002). Recently, the variable P6b region has also been used for constructing phylogenies (Linke *et al.*, 2003; Wirtz *et al.*, 2003). However, in our phylogenetic analyses of the 16S rRNA gene and of 16S rRNA gene and trnL intron combined, some cyanobacteria sharing identical repeat classes in their trnL intron P6b loop fell into paraphyletic groups consistently (Fig. 2a). Thus, the analyses of the 16S rRNA gene sequences show that cyanobacterial strains with the same repeat class in their trnL intron are not always more closely related to strains with a different repeat class or without these repeats. For example, three *Nostoc* strains that harboured the class

I repeat in their P6b intron loop were intermixed with strains that harboured the class II repeat (Fig. 2a). *Nostoc* 9.4.19 with class I repeat and *Nostoc* 9.4.22 with class II repeat were monophyletic in phylogenetic analyses of the conserved trnL intron sequence but highly different in their P6b sequence (Fig. 2). While, in the P6b sequence, they had 39 nt (49 %) differences, in all other parts of the intron, the difference was only 3 nt, and sequence similarity of the 16S rRNA gene was 98 %. Similarly, strains *Calothrix* PCC 7102 with class I repeat and *Calothrix* PCC 7714 with class II repeat had 31 nt differences (including 10 gaps) in their P6b sequence but they were monophyletic in the 16S rRNA gene tree, indicating that nucleotide differences in the P6b loop do not predict strain phylogeny, and therefore it is essential to exclude the stem-loop in phylogenetic analyses. The Bayesian taxon partition on a much larger number of trees (hypotheses) than in the ML analysis also supported this finding. The group of *Nostoc* strains with both repeat classes (*Nostoc* B in Fig. 2a) was monophyletic with full Bayesian posterior probability in every tree. Finally, the alternative hypothesis of a monophyletic origin of the *Nostoc* strains containing class I repeats in the P6b loop of the trnL intron was rejected by a statistically highly significant ($P=0.001$) difference in the Shimodaira–Hasegawa test (Shimodaira & Hasegawa, 1999). Hence, it is highly unlikely that the 16S rRNA gene tree topology in Fig. 2(a) would have been obtained by chance. However, although the 16S rRNA gene is a common phylogenetic tool (Woese, 2000) and the test supported our 16S rRNA tree, we cannot explicitly trust the 16S rRNA gene phylogeny alone without knowledge of possible variation among 16S rRNA gene copies. Nevertheless, phylogenetic analyses of the conserved intron sequences without the most variable loops resulted in a tree topology similar to that of the 16S rRNA gene tree, thus providing additional support for the 16S rRNA gene phylogeny.

The polyphyly of the repeat classes indicates that the evolutionary history of the two repeat classes in the P6b loops of the trnL intron of heterocystous cyanobacteria differs from the strain phylogeny as indicated by the 16S rRNA gene and the conserved parts of the trnL intron sequences. Thus, similar P6b sequences may lead to wrong phylogenies. They should not be taken as reliable evidence of phylogenetic relationships and the stem-loop should not be included in the phylogenetic analyses. This might explain why the 16S rRNA gene tree and the trnL intron tree of cyanobacteria were incongruent in a previous study that included parts of the intron stem-loops in the phylogenetic analyses (Rudi & Jakobsen, 1999). Another explanation could be that the 485-nt long alignment of 16S rRNA gene sequences with too few informative characters can result in erroneous or poorly resolved phylogeny (parsimony analysis; data not shown). Wirtz *et al.* (2003) used another approach by excluding sequence fragments from the beginning and end of the trnL intron stem region and including in analyses the total P6b loop sequence that contained both nucleotide and length variation due to

Fig. 2. Bayesian majority rule (50%) consensus trees showing phylogenetic relationships among cyanobacteria based on sequence data of the 16S rRNA gene (a), conserved parts of trnL intron (b) and conserved parts of trnL intron with ingroup sampling (c). Bootstrap support proportions 50–100 (%) and Bayesian posterior probabilities of 0.8–1.0 are shown at nodes. The distribution of trnL intron repeat class I is shown in the 16S rRNA gene tree as black circles and class II repeats as white squares. Underlined strains with class I repeats indicate the polyphyletic distribution of the repeats. Intron length is given as bp in (b).

repeats. A repeat, however, is seen as a unit of one mutational event (e.g. Costa *et al.*, 2002). Therefore, if the vertical inheritance and homology of the sequences has been confirmed, coding of single events and removing nucleotide sequence of repeats from the alignment would be a more preferable method. Nevertheless, as we show here that repeats in the trnL intron can lead to erroneous phylogenies, they should not be used at all in phylogenetic analyses.

In the case of the trnL intron sequence, a small number of informative characters (Table 1) clearly restricts phylogenetic group formation. This relates to the short length of the sequences and to the stable secondary and tertiary structure required by the self-splicing property of the group I intron (Xu *et al.*, 1990), which restricts the number of random mutations in the stem regions. In the study of Paquin *et al.* (1997), a subtree based mainly on the intron stem was highly unresolved. In our analyses, the inclusion of cyanobacteria that were distantly related to *Nostoc* also resulted in unresolved trees (not shown). Only after the distant relatives were removed, did some hierarchy become evident (Fig. 2b, c), although only with weak support values. Hence, the conserved parts of the trnL intron alone do not provide enough sequence variation for hierarchical analyses. As the conserved intron sequence was combined with the 16S rRNA gene data, the resulting tree topology in main clades (Supplementary Figure) was nearly identical to the 16S rRNA gene tree without the intron, indicating that phylogenetic information of the conserved parts of the intron sequence was not in strong conflict with that of the 16S rRNA gene. We did not obtain trnL intron PCR products from some strains of the genera *Anabaena*, *Aphanizomenon*, *Calothrix*, *Cyanothece*, *Fisherella*, *Microcystis*, *Oscillatoria* and *Tolypothrix* (Supplementary Table B). The missing introns resulted either from technical problems or from sporadic distribution of the intron. It is obvious that sporadic distribution of the intron will be problematic if it is used as a molecular marker among strains that potentially lack it.

Congruence of phylogenetic methods

In accordance with Suzuki *et al.* (2002) and Douady *et al.* (2003), the bootstrap support values and Bayesian posterior probabilities were not always in the same category (Fig. 2). However, the phylogenetic trees obtained by different phylogenetic methods were generally very similar, with only some minor differences in tree topology. *Nostoc* group A (Fig. 2a) was well supported in all analyses, but *Nostoc* strain VI.2.2 fell outside this well-resolved group in the Bayesian analyses of 16S rRNA gene sequences and the combined dataset. Moreover, the weak posterior probability of this branch in Bayesian analyses indicated that the position of the strain was unresolved. In a Bayesian taxon partition with the 16S rRNA gene data, the monophyly of *Nostoc* A (without strain VI.2.2) was found in separate runs in nearly all trees.

Distribution and hypothetical origin of the repeats

The length of trnL intron sequences varied greatly, ranging from 257 to 346 nt (Fig. 2b; Supplementary Table A) due to the loops, mainly P6b and P9. Generally, *Nostoc* strains had long introns (272–346 nt), usually with four repeat copies and sometimes with additional insertions in their P6b loops. The trnL introns from *Pseudanabaena*, *Synechococcus* and *Aphanizomenon* strains were the shortest and only had a 17-nt P6b. While only the *Nostoc* and the *Calothrix* strains had recognizable repeat classes in their P6b loops, *Nodularia* also had one class II heptanucleotide motif and *Cylindrospermum* two motifs close to the class I repeat. The P5 loops of the two *Anabaena* strains were six nucleotides longer than any other strain, while *Pseudanabaena* and *Planktothrix* had the shortest, these being only three and four nucleotides long. Large length variation in loops demonstrates how difficult it is to find evolutionary homologies between single nucleotides in intron loops. The class II repeat sequence GCTGAGT of the trnL intron of some *Nostoc* strains, a *Nodularia* strain and a *Tolypothrix* strain was found as two copies in an internal transcribed spacer between tRNA^{Ala} and 23S genes (rRNA operon) from a heterocystous *Gloeotrichia* species (AF105135). The sequence TGCTGAGT AATGAGT GCTGAGT GTTGAGT in the P6b loop of the trnL intron of *Nostoc* 152 (PCC 9237/1) was found with two mismatches and an E-value of 0.01 from the *mcvB* gene of *Nostoc* PCC 7120 (total genome available at <http://www.kazusa.or.jp/cyano/blast.html>) and with one mismatch in the internal transcribed spacer of the *Gloeotrichia* species (underlined nucleotides). The 26-nt AT-rich repeat site in the P6b loops of the introns of *Nostoc* strains, AAAATTCAAAATCT AAAATCCAAAAT was found with one mismatch from other heterocystous cyanobacteria. This sequence was found in the repeat family STRR6 between the *nifS* and *nifU* genes of *Anabaena azollae* (Jackman & Mulligan, 1995), between the genes *rbcL* and *rbcX* in *Anabaena* 133 (Gugger *et al.*, 2002) and in the repeat region of the *apcEA1B1C* operon of *Calothrix* PCC 7601 (Houmard *et al.*, 1990). As the class II repeat of the trnL intron and other tandemly repeated heptanucleotide sequences (Meeks *et al.*, 2001) are found at multiple sites in many heterocystous cyanobacteria, they are not trnL intron specific and seem to lack the functional constancy and genetic stability required from phylogenetic markers (Ludwig & Schleifer, 1999).

The tandemly repeated heptamers in the P6b loop of the trnL intron correspond to a common definition of mini-satellites by size (6- to 100-nt motif spanning 0.5 kb to several kilobases) and by being hypervariable in the loci (Vergnaud & Denoeud, 2000). The evolutionary origin of the heptanucleotide repeat classes is unclear. Possible mechanisms, such as slipped-strand mispairing and errors of DNA polymerase in DNA duplication, have been suggested (Costa *et al.*, 2002). Repeats may also originate from intra- or intergenomic recombination, from transposase or they may have been transferred by a phage

(Lewin, 2000). Theories of lateral intron transfer and recombination of the entire tRNA locus suggested for the trnL intron type II (Rudi & Jakobsen, 1997; Rudi *et al.*, 2002) may not explain the evolution of intron type I, since cyanobacteria with different repeat classes in their introns nonetheless shared the most closely related introns as indicated by the conserved stem sequences of the introns. Thus, the entire locus may not have been transferred laterally, but only the P6b region of the intron may have undergone recombination between strains. Among *Nostoc* strains, various evolutionary events have occurred in the P6b region. Although most *Nostoc* strains here have long P6b sequences, with repeats and other insertions, some *Nostoc commune* strains lack the P6b loop (Wright *et al.*, 2001). It would be necessary to confirm their genus identity and phylogeny to reveal the most probable evolutionary explanation for P6b in those strains.

To conclude, sequence differences in the hypervariable regions of the trnL intron cannot be used as a measure of taxonomic relationships. In contrast to eukaryotic micro- and minisatellites, repeats in the trnL intron are not applicable in population genetics, since their evolution seems to involve not only verticality but also laterality.

ACKNOWLEDGEMENTS

This study was financially supported by the Academy of Finland (projects 168332 and 202162). Mikko Kolkkala and Dr Riitta Savolainen are thanked for giving I.O. valuable advice about MrBayes. Mikko Kolkkala, Dr Christina Lyra and Dr David Fewer are acknowledged for critical comments on the manuscript.

REFERENCES

- Alfaro, M. E., Zoller, S. & Lutzoni, F. (2003). Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Mol Biol Evol* **20**, 255–266.
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.
- Besendahl, A., Qiu, Y.-L., Lee, J., Palmer, J. D. & Bhattacharya, D. (2000). The cyanobacterial origin and vertical transmission of the plastid tRNA^{Leu} group-I intron. *Curr Genet* **37**, 12–23.
- Biniszkievicz, D., Cesnaviciene, E. & Shub, D. A. (1994). Self-splicing group I intron in cyanobacterial initiator methionine tRNA: evidence for lateral transfer of introns in bacteria. *EMBO J* **13**, 4629–4635.
- Castenholz, R. W. (2001). Phylum BX. Cyanobacteria. In *Bergey's Manual of Systematic Bacteriology*, 2nd edn, vol. 1, pp. 473–599. Edited by D. R. Boone & R. W. Castenholz. New York: Springer.
- Cech, T. R., Damberger, S. H. & Gutell, R. R. (1994). Representation of the secondary and tertiary structures of group I introns. *Nat Struct Biol* **1**, 273–280.
- Costa, J. L., Paulsrud, P. & Lindblad, P. (1999). Cyanobiont diversity within coraloid roots of selected cycad species. *FEMS Microbiol Ecol* **28**, 85–91.
- Costa, J. L., Paulsrud, P., Rikkinen, J. & Lindblad, P. (2001). Genetic diversity of *Nostoc* symbionts endophytically associated with two bryophyte species. *Appl Environ Microbiol* **67**, 4393–4396.
- Costa, J. L., Paulsrud, P. & Lindblad, P. (2002). The cyanobacterial tRNA^{Leu} (UAA) intron: evolutionary patterns in a genetic marker. *Mol Biol Evol* **19**, 850–857.
- Douady, C. J., Delsuc, F., Boucher, Y., Doolittle, W. F. & Douzery, E. J. P. (2003). Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol Biol Evol* **20**, 248–254.
- Goldman, N., Anderson, J. P. & Rodrigo, A. G. (2000). Likelihood-based tests of topologies in phylogenetics. *Syst Biol* **49**, 652–670.
- Gugger, M., Lyra, C., Henriksen, P., Coute, A., Humbert, J. F. & Sivonen, K. (2002). Phylogenetic comparison of the cyanobacterial genera *Anabaena* and *Aphanizomenon*. *Int J Syst Evol Microbiol* **52**, 1867–1880.
- Houmard, J., Capuano, V., Colombano, M. V., Coursin, T. & Tandeau de Marsac, N. (1990). Molecular characterization of the terminal energy acceptor of cyanobacterial phycobilisomes. *Proc Natl Acad Sci U S A* **87**, 2152–2156.
- Huelsenbeck, J. P. (2000). MrBayes: Bayesian inference of phylogeny. Distributed by the author. Department of Biology, University of Rochester, NY, USA.
- Huelsenbeck, J. P. & Ronquist, F. (2001). MrBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755.
- Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310–2314.
- Ishida, T., Watanabe, M. M., Sugiyama, J. & Yokota, A. (2001). Evidence for polyphyletic origin of the members of the orders of Oscillatoriales and Pleurocapsales as determined by 16S rDNA analysis. *FEMS Microbiol Lett* **201**, 79–82.
- Jackman, D. M. & Mulligan, M. E. (1995). Characterization of a nitrogen-fixation (*nif*) gene cluster from *Anabaena azollae* 1a shows that closely related cyanobacteria have highly variable but structured intergenic regions. *Microbiology* **141**, 2235–2244.
- Kuhse, M. G., Strickland, R. & Palmer, J. D. (1990). An ancient group I intron shared by eubacteria and chloroplasts. *Science* **250**, 1570–1573.
- Lehtimäki, J., Lyra, C., Suomalainen, S., Sundman, P., Rouhiainen, L., Paulin, L., Salkinoja-Salonen, M. & Sivonen, K. (2000). Characterization of *Nodularia* strains, cyanobacteria from brackish waters, by genotypic and phenotypic methods. *Int J Syst Evol Microbiol* **50**, 1043–1053.
- Lewin, B. (2000). *Genes VII*, part 4, pp. 347–479. Oxford: Oxford University Press.
- Linke, K., Hemmerich, J. & Lumbsch, H. T. (2003). Identification of *Nostoc* cyanobionts in some *Peltigera* species using a group I intron in the tRNA^{Leu} gene. *Bibliotheca Lichenol* **86**, 113–118.
- Lohtander, K., Källersjö, M., Moberg, R. & Tehler, A. (2000). The family Physciaceae in Fennoscandia: phylogeny inferred from ITS sequences. *Mycologia* **92**, 728–735.
- Ludwig, W. & Schleifer, K. H. (1999). Phylogeny of Bacteria beyond the 16S rRNA standard. *ASM News* **65**, 752–757.
- Lyra, C., Suomalainen, S., Gugger, M., Vezie, C., Sundman, P., Paulin, L. & Sivonen, K. (2001). Molecular characterization of planktic cyanobacteria of *Anabaena*, *Aphanizomenon*, *Microcystis* and *Planktothrix* genera. *Int J Syst Evol Microbiol* **51**, 513–526.
- Meeks, J. C., Elhai, J., Thiel, T., Potts, M., Larimer, F., Lamerdin, J., Predki, P. & Atlas, R. (2001). An overview of the genome of *Nostoc*

punctiforme, a multicellular, symbiotic cyanobacterium. *Photosynth Res* **70**, 85–106.

Miao, V. P. W., Rabenau, A. & Lee, A. (1997). Cultural and molecular characterization of photobionts of *Peltigera membranacea*. *Lichenologist* **29**, 571–586.

Oksanen, I., Lohtander, K., Paulsruud, P. & Rikkinen, J. (2002). A molecular approach to cyanobacterial diversity in a rock-pool community involving gelatinous lichens and free-living *Nostoc* colonies. *Ann Bot Fennici* **39**, 93–99.

Page, R. D. M. (1996). TREEVIEW: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* **12**, 357–358.

Paquin, B., Kathe, S. D., Nierzwicki-Bauer, S. A. & Shub, D. A. (1997). Origin and evolution of group-I introns in cyanobacterial tRNA genes. *J Bacteriol* **179**, 6798–6806.

Paulsruud, P. (2001). The *Nostoc* symbiont of lichens: diversity, specificity and cellular modifications. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. *Acta Univ Upsaliensis* **662**, 1–55.

Paulsruud, P. & Lindblad, P. (1998). Sequence variation of the tRNA^{Leu} intron as a marker for genetic diversity and specificity of symbiotic cyanobacteria in some lichens. *Appl Environ Microbiol* **64**, 310–315.

Paulsruud, P., Rikkinen, J. & Lindblad, P. (1998). Cyanobiont specificity in some *Nostoc*-containing lichens and a *Peltigera aphthosa* photosymbiodeme. *New Phytol* **139**, 517–524.

Paulsruud, P., Rikkinen, J. & Lindblad, P. (2000). Spatial patterns of photobiont diversity in some *Nostoc*-containing lichens. *New Phytol* **146**, 291–299.

Paulsruud, P., Rikkinen, J. & Lindblad, P. (2001). Field investigations on cyanobacterial specificity in *Peltigera aphthosa*. *New Phytol* **152**, 117–123.

Posada, D. & Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818.

Posada, D. & Crandall, K. A. (2001). Selecting the best-fit model of nucleotide substitution. *Syst Biol* **50**, 580–601.

Rikkinen, J., Oksanen, I. & Lohtander, K. (2002). Ecological lichen guilds share related cyanobacterial symbionts. *Science* **297**, 357.

Rudi, K. & Jakobsen, K. S. (1997). Cyanobacterial tRNA^{Leu} (UAA) group-I introns have a polyphyletic origin. *FEMS Microbiol Lett* **156**, 293–298.

Rudi, K. & Jakobsen, K. S. (1999). Complex evolutionary patterns of tRNA^{Leu} (UAA) group-I introns in cyanobacterial radiation. *J Bacteriol* **181**, 3445–3451.

Rudi, K., Fossheim, T. & Jakobsen, K. S. (2002). Nested evolution of a tRNA^{Leu} (UAA) group I intron by both horizontal intron transfer and recombination of the entire tRNA locus. *J Bacteriol* **184**, 666–671.

Shimodaira, H. & Hasegawa, M. (1999). Multiple comparison of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* **16**, 1114–1116.

Strehl, B., Holtzendorff, J., Partensky, F. & Hess, W. R. (1999). A small and compact genome in the marine cyanobacterium *Prochlorococcus marinus* CCMP 1375: lack of an intron in the gene for tRNA^{Leu} (UAA) and a single copy of the rRNA operon. *FEMS Microbiol Lett* **181**, 261–266.

Summerfield, T. C., Galloway, D. J. & Eaton-Rye, J. J. (2002). Species of cyanolichens from *Pseudocyphellaria* with indistinguishable ITS sequences have different photobionts. *New Phytol* **155**, 121–129.

Suzuki, Y., Glazko, G. V. & Nei, M. (2002). Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc Natl Acad Sci U S A* **99**, 16138–16143.

Swofford, D. L. (2002). PAUP*: Phylogenetic analysis using parsimony (* and other methods), version 4. Sunderland, MA: Sinauer Associates.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680.

Turner, S., Pryer, K. M., Miao, V. P. W. & Palmer, J. D. (1999). Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *J Eukaryot Microbiol* **46**, 327–338.

Vepritskiy, A. A., Vitol, I. A. & Nierzwicki-Bauer, S. A. (2002). Novel group I intron in the tRNA^{Leu} (UAA) gene of a γ -proteobacterium isolated from a deep subsurface environment. *J Bacteriol* **184**, 1481–1487.

Vergnaud, G. & Denoeud, F. (2000). Minisatellites: mutability and genome architecture. *Genome Res* **10**, 899–907.

Wilmotte, A., Van der Auwera, G. & De Wachter, R. (1993). Structure of the 16S ribosomal RNA of the thermophilic cyanobacterium *Chlorogloeopsis* HTF ('*Mastigocladus laminosus* HTF') strain PCC7518, and phylogenetic analysis. *FEBS Lett* **317**, 96–100.

Wirtz, N., Lumbsch, H. T., Green, T. G. A., Türk, R., Pintado, A., Sancho, L. & Schroeter, B. (2003). Lichen fungi have low cyanobiont selectivity in maritime Antarctica. *New Phytol* **160**, 177–183.

Woese, C. R. (2000). Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci U S A* **97**, 8392–8396.

Wright, D., Prickett, T., Helm, R. F. & Potts, M. (2001). Form species *Nostoc commune* (Cyanobacteria). *Int J Syst Evol Microbiol* **51**, 1839–1852.

Xu, M.-Q., Kathe, S. D., Goodrich-Blair, H., Nierzwicki-Bauer, S. A. & Shub, D. A. (1990). Bacterial origin of a chloroplast intron: conserved self-splicing group I introns in cyanobacteria. *Science* **250**, 1566–1570.