Evolution and variation of the nifD and hupL elements in the heterocystous cyanobacteria

Brian J. Henson,1 Linda Hartman,2 Linda E. Watson3 and Susan R. Barnum2

1Department of Pediatrics, Children’s Hospital of Pittsburgh of UPMC, USA
2Department of Botany, Miami University, USA
3Department of Botany, Oklahoma State University, USA

In heterocystous cyanobacteria, heterocyst differentiation is accompanied by developmentally regulated DNA rearrangements that occur within the nifD and hupL genes, referred to as the nifD and hupL elements. These elements are segments of DNA that are embedded within the coding region of each gene and range from 4 to 24 kb in length. The nifD and hupL elements are independently excised from the genome during the later stages of differentiation by the site-specific recombinases, XisA and XisC, respectively, which are encoded within the elements themselves. Here we examine the variation and evolution of the nifD and hupL elements by comparing full-length nifD and hupL element sequences and by phylogenetic analysis of xisA and xisC gene sequences. There is considerable variation in the size and composition of the nifD and hupL elements, however, conserved regions are also present within representatives of each element. The data suggest that the nifD and hupL elements have undergone a complex pattern of insertions, deletions, translocations and sequence divergence over the course of evolution, but that conserved regions remain.

INTRODUCTION

The heterocystous cyanobacteria are unique in that they develop heterocysts, which are specialized cells with thickened cell walls that prevent nitrogenase from coming into contact with oxygen, thereby spatially separating photosynthesis from nitrogen fixation (Adams & Duggan, 1999; Böhme, 1998; Wolk et al., 1994). Heterocysts supply adjacent vegetative cells with reduced nitrogen and, in return, the vegetative cells supply the heterocysts with photosynthates (Adams & Duggan, 1999; Lynn et al., 1986). Heterocyst differentiation is a complex process involving up to 1000 different genes, or approximately 15–25% of the cyanobacterial genome (Adams & Duggan, 1999; Böhme, 1998; Wolk et al., 1994). Heterocyst differentiation is accompanied by a developmental shift or differentiation process. Cellular differentiation and developmentally regulated DNA rearrangements are uncommon in prokaryotes, with the best studied examples being sporulation in Bacillus subtilis and Clostridium difficile, as well as heterocyst differentiation in cyanobacteria (Kunkel et al., 1990; Haselkorn, 1992; Haraldsen & Sonenshein, 2003).

In cyanobacteria, DNA rearrangements occur within the nifD, fixN and hupL genes, which encode a subunit of nitrogenase, the ferredoxin protein and a subunit of the membrane-bound uptake hydrogenase, respectively (Golden et al., 1985, 1987, 1991; Mulligan et al., 1988; Oxelfelt et al., 1985; Haselkorn, 1992; Carrasco & Golden, 1995; Carrasco et al., 1995; Wolk et al., 1994). In each of these three genes, specific segments of DNA, hereafter referred to as DNA elements, are independently excised from the genome during the later stages of heterocyst differentiation by site-specific recombination between direct repeats that flank each element (Adams & Duggan, 1999; Haselkorn, 1992; Carrasco et al., 1995). The length and nucleotide sequence of the direct repeats varies for each element type (Adams & Duggan, 1999) and each element encodes its own site-specific recombinase responsible for its excision (Adams & Duggan, 1999; Böhme, 1998). All three DNA elements have been found in Nostoc sp. strain PCC 7120 (Golden et al., 1985, 1987, 1991), but otherwise little is known about their co-occurrence and distribution within the heterocystous cyanobacteria.

The three types of DNA elements vary in their size and pattern of excision. An 11 kb element is present within the nifD gene of Nostoc sp. strain PCC 7120. The NifD gene is
part of the \textit{nifHDK} operon, which encodes nitrogenase (Mazur & Chui, 1982). Expression of sequences downstream of the element in the \textit{nifHDK} operon and nitrogen fixation are contingent on the removal of this element. It is excised between 11 bp direct repeats (CGGAATTCCT) that flank the element by XisA, which is encoded within the element (Brusca \textit{et al.}, 1989, 1990; Golden \textit{et al.}, 1985, 1987, 1991; Lammers \textit{et al.}, 1986, 1990; Haselkorn, 1992). Many other potential open reading frames (ORFs) are located within the element; although it is not believed that any gene other than \textit{xisA} is required for excision (Lammers \textit{et al.}, 1986, 1990; Brusca \textit{et al.}, 1990). A 55 kb element is embedded within the \textit{fxdN} gene of \textit{Nostoc} sp. strain PCC 7120. The \textit{fxdN} gene is part of the \textit{nifBSU}, \textit{fxdN} operon (Golden \textit{et al.}, 1987, 1988; Mulligan \textit{et al.}, 1988; Carrasco \textit{et al.}, 1994) and removal of the element is required for expression of sequences downstream of the element. The element is flanked by 5 bp direct repeats (TATTC) and it is excised by XisA and its accessory proteins XisH and XisI, all of which are encoded within the \textit{fxdN} element (Ramaseswamy \textit{et al.}, 1997). Embedded within the \textit{hupL} gene of \textit{Nostoc} sp. strain PCC 7120 is a 9.4 kb element (Carrasco \textit{et al.}, 1995, 2005). This element is flanked by 16 bp direct repeats (CCATATAACTGCTGTG) and it is excised by the recombinase XisC, which is encoded within the element (Carrasco \textit{et al.}, 1995, 2005) (Fig. 2). It is believed that no other functional products are encoded within the \textit{hupL} element (Carrasco \textit{et al.}, 1995, 2005).

Although these programmed DNA rearrangements have received moderate attention in the literature, many questions remain regarding their evolution, variation and regulation of excision. It is not known how widespread these elements are in heterocystous cyanobacteria, however the \textit{nifD} and \textit{hupL} elements have been found in a number of strains (Tables 1 and 2, respectively), whereas the \textit{fxdN} element has only been found in three strains: \textit{Nostoc} sp. strain Mac, \textit{Anabaena} sp. strain PCC 7122 and \textit{Anabaena} sp. strain M131 (Carrasco & Golden, 1995). It is generally believed that only heterocystous cyanobacteria of Subsection IV harbour the elements (Saville \textit{et al.}, 1987; Kallas \textit{et al.}, 1985; Lammers \textit{et al.}, 1990); however, a recent study reported the presence of the \textit{hupL} element (or a portion of \textit{xisC}) in \textit{Fischlerella} sp. strain PCC 7521 (Subsection V) and \textit{Leptolyngbya} sp. strain PCC 73110 (a non-heterocystous Subsection III strain) (Table 2) (Tamagnini \textit{et al.}, 2000). Within Subsection IV, the distribution of the \textit{nifD} and \textit{hupL} elements appears to be haphazard and their pattern of occurrence does not appear to be linked (Happe \textit{et al.}, 2000; Meeks \textit{et al.}, 2001).

The origins of these DNA elements remain unknown. It has been suggested that the \textit{nifD}, \textit{hupL} and \textit{fxdN} elements are the remnants of ancient viral infections and have lost the ability to self-replicate (Haselkorn, 1992; Henson \textit{et al.}, 2005). The \textit{fxdN} element has sequence similarity to the \textit{Bacillus subtilis} \textit{SKIN} element (48 kb) that is excised from within the \textit{sigK} gene during sporulation by the SpoIVCA recombinase (Kunkel \textit{et al.}, 1990; Haselkorn, 1992; Carrasco \textit{et al.}, 1994). Both XisF and SpoIVCA are members of the resolvase/invertase or serine family of recombinases (Smith & Thorpe, 2002). XisA and XisC belong to the tyrosine or phage integrase family of site-specific recombinases (Nunes-Düby \textit{et al.}, 1998). The tyrosine integrases are characterized by a conserved R-H-Y tetrad within the active site (Voziyano\textit{v} et al., 1999; Nunes-Düby \textit{et al.}, 1998); however, XisA and XisC have an altered active site with a tyrosine as opposed to a histidine (R-Y-R-Y) (Nunes-Düby \textit{et al.}, 1998). XisA and XisC are 61 % similar at the nucleotide level and 43 % similar at the amino acid level (Carrasco \textit{et al.}, 1995). BLAST (basic local alignment search tool) searches indicate that they are more similar to each other than to any other sequence in GenBank. The similarities between \textit{xisA} and \textit{xisC} suggest that they may share a common ancestor and this also suggests that the \textit{nifD} and \textit{hupL} elements themselves may share a common evolutionary origin and may have arisen by subsequent duplication and divergence prior to their insertion within the cyanobacterial lineage.

Our objectives were to examine the structural and compositional variation within the \textit{nifD} and \textit{hupL} elements and examine whether or not they are of viral origin. This was accomplished by comparative analysis of full-length \textit{nifD} and \textit{hupL} gene sequences and phylogenetic analysis of \textit{xisA} and \textit{xisC} gene sequences.

**METHODS**

Cyanobacterial strains (Table 3) were obtained from the Pasteur Culture Collection (PCC) and American type Culture Collection (ATCC) and were grown in an illuminated shaking incubator at 27 °C in a variation of BG-11 media [BG-110 + NaHCO\textsubscript{3} (5 mM)], which contained reduced nitrate. Genomic DNA was extracted using a Puregene DNA Isolation kit (Genta Systems) with a slight modification (Henson \textit{et al.}, 2002).

Amplification of full-length \textit{nifD} and \textit{hupL} elements was accomplished with the Expand Long Template PCR system (Roche). For the \textit{nifD} element, the primers used for PCR amplification were 5'-TTTCCGGCAAATGCACTCTTG-3' and 5'-GCAAAACGTGTGATACCGTGAT-3', which are located within the flanking regions of the \textit{nifD} direct repeats. To amplify the \textit{hupL} element, the primers used for PCR amplification were 5'-GCGCGCATACACACCAAGCC-3' and 5'-GCGCTGTGGTAGCGCAATGGG-3', which are located within the flanking regions of the \textit{hupL} direct repeats. PCR conditions for amplifying full-length elements were as follows: an initial hold at 94 °C for 2 min, 10 cycles of denaturation at 94 °C for 10 s, annealing at 55 °C for 30 s, and extension at 68 °C for 10 min followed by 15 cycles of denaturation at 94 °C for 10 s, annealing at 55 °C for 30 s, extension at 68 °C for 10 min, with each cycle increasing 20 s in duration, and a final extension of 68 °C for 7 min. PCR products were verified by gel electrophoresis on 0.8% agarose and purified with the QiAquick Gel Extraction kit (Qiagen). The cyanobacterial strains from which \textit{nifD} and \textit{hupL} elements were sequenced are listed in Tables 1 and 2, respectively.

Amplification of partial \textit{xisA} and \textit{xisC} genes was accomplished with \textit{Taq} DNA polymerase (New England Biolabs). The cyanobacterial strains from which partial \textit{xisA} and \textit{xisC} sequences were generated are listed in Tables 1 and 2, respectively. Primers used to amplify \textit{xisA} were 5'-CTGCCCTTGAAGTCTGCAAAGAC-3' and 5'-GCTTAAATCGTCTCAGGCCTT-3'.
Table 1. Cyanobacteria examined for the presence of the nifD element

Strains are separated by taxonomic Subsection. Presence refers to whether or not the element has been detected. Method refers to whether the element was detected by Southern hybridization (S) or by DNA sequencing (D). Y, Yes; N, no; NA, not available; C, completely sequenced.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Presence</th>
<th>Method</th>
<th>Length sequenced</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subsection I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aphanothece sp. ATCC 27178</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Saville et al. (1987)</td>
</tr>
<tr>
<td>Anacystis nidulans R2</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Saville et al. (1987)</td>
</tr>
<tr>
<td>Coccochloris peniocystis</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Saville et al. (1987)</td>
</tr>
<tr>
<td>Cyanothecae sp. PCC 7424</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Kallas et al. (1985)</td>
</tr>
<tr>
<td>Gloeocapsa sp. ATCC 27191</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Saville et al. (1987)</td>
</tr>
<tr>
<td>Synechococcus cedrorum</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Saville et al. (1987)</td>
</tr>
<tr>
<td>Synechococcus sp. PCC 6301</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Kallas et al. (1985)</td>
</tr>
<tr>
<td>Synechococcus sp. PCC 7335</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Kallas et al. (1985)</td>
</tr>
<tr>
<td>Synechococcus sp. PCC 7425</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Kallas et al. (1985)</td>
</tr>
<tr>
<td>Synechocystis sp. ATCC 29109</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Saville et al. (1987)</td>
</tr>
<tr>
<td>Subsection III</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leptolyngbya sp. ATCC 29126</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Saville et al. (1987)</td>
</tr>
<tr>
<td>Leptolyngbya sp. PCC 73110</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Kallas et al. (1985)</td>
</tr>
<tr>
<td>Pseudoanabaena sp. ATCC 29210</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Saville et al. (1987)</td>
</tr>
<tr>
<td>Pseudoanabaena sp. PCC 7409</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Kallas et al. (1985)</td>
</tr>
<tr>
<td>Subsection V</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fischerella sp. ATCC 27929</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Saville et al. (1987)</td>
</tr>
<tr>
<td>Mastigocladus sp. ARM 351</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Prasanna &amp; Kaushik (1995)</td>
</tr>
<tr>
<td>Subsection IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anabaena azollae</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Meeks et al. (1988)</td>
</tr>
<tr>
<td>Anabaena sp. ATCC 33047</td>
<td>Y</td>
<td>D</td>
<td>NA</td>
<td>Henson et al. (2005)</td>
</tr>
<tr>
<td>Anabaena sp. PCC 7108</td>
<td>Y</td>
<td>D</td>
<td>xisA (700 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Anabaena. variabilis</td>
<td>Y</td>
<td>D</td>
<td>xisA (700 bp)</td>
<td>Brusca et al. (1989)</td>
</tr>
<tr>
<td>Calothrix sp. PCC 7601</td>
<td>Y</td>
<td>S</td>
<td>NA</td>
<td>Kallas et al. (1985)</td>
</tr>
<tr>
<td>Cyanothece sp. PCC 7604</td>
<td>Y</td>
<td>D</td>
<td>xisA (500 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Cyanothece sp. PCC 71417</td>
<td>Y</td>
<td>D</td>
<td>C</td>
<td>This study</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7210</td>
<td>Y</td>
<td>D</td>
<td>C</td>
<td>Kaneko et al. (2001)</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7121</td>
<td>Y</td>
<td>D</td>
<td>xisA (700 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7906</td>
<td>Y</td>
<td>D</td>
<td>xisA (500 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Tolypothrix sp. PCC 7415</td>
<td>Y</td>
<td>D</td>
<td>xisA (700 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Nostoc sp. PCC 6314</td>
<td>Y</td>
<td>D</td>
<td>xisA (700 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7118</td>
<td>Y</td>
<td>D</td>
<td>xisA (700 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7416</td>
<td>Y</td>
<td>D</td>
<td>xisA (700 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7524</td>
<td>Y</td>
<td>D</td>
<td>xisA (700 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Nostoc punctiforme</td>
<td>Y</td>
<td>D</td>
<td>C</td>
<td>Meeks et al. (2001)</td>
</tr>
<tr>
<td>Nostoc sp. strain Mac</td>
<td>Y</td>
<td>S</td>
<td>NA</td>
<td>Meeks et al. (1994)</td>
</tr>
<tr>
<td>Scytonema sp. PCC 7110</td>
<td>Y</td>
<td>D</td>
<td>xisA (700 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Scytonema sp. ARM 428</td>
<td>Y</td>
<td>S</td>
<td>none</td>
<td>Prasanna &amp; Kaushik (1995)</td>
</tr>
<tr>
<td>Synechococcus sp. ARM 428</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Prasanna &amp; Kaushik (1995)</td>
</tr>
<tr>
<td>Tolypothrix ceylonica ARM397</td>
<td>Y</td>
<td>S</td>
<td>none</td>
<td>Prasanna &amp; Kaushik (1995)</td>
</tr>
</tbody>
</table>

CAAAAACTGCTAAACACTCC-3', To amplify xisC, the primers 5'-GGAGACTCCAACCGACGTGGTGCGC-3' and 5'-TATGATTGTC-TCTAAGGATACA-3' were used. PCR conditions included: an initial hold at 94 °C for 2 min, 35 cycles of denaturation at 94 °C for 20 s, annealing at 55 °C for 15 s, and extension at 72 °C for 1 min. PCR products were verified by gel electrophoresis on 0.8% agarose and were purified with a QIAquick Gel Extraction kit (Qiagen).

Sequencing of purified PCR products was accomplished with a DYEnamic ET Terminator cycle sequencing kit (Amersham Biosciences) using capillary electrophoresis on ABI 310, ABI 3100 and ABI 3730 Genetic Analyzers (Applied Biosystems). Both the forward and reverse DNA strands of the PCR products (full-length nifD and hupL elements and partial xisA and xisC genes) were sequenced to ensure that the nucleotide sequences were accurate and that no artefacts of PCR were present. Sequences were deposited in GenBank and the accession numbers are listed in Table 3.

BLAST searches were carried out using complete nifD and hupL element sequences, as well as smaller portions of each element (ranging from ~200 to 1500 bp) to determine regions of similarity with other sequences in the database. The BLAST programs used were: quick search for highly similar sequences (MEGABLAST), quick search for divergent sequences (discontiguous MEGABLAST), nucleotide–nucleotide BLAST (BLASTN), search for short, nearly exact matches and translated query vs protein database (BLASTX). Databases that
Table 2. Cyanobacterial strains that have been examined for the presence of the hupL element

Strains are separated by Subsection. Presence refers to whether or not the element has been detected. Method refers to whether the element was detected by Southern hybridization (S) or by DNA sequencing (D). Y, Yes; N, no; NA, not available; C, completely sequenced.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Presence</th>
<th>Method</th>
<th>Length sequenced</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subsection III</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leptolyngbya sp. PCC 73110</td>
<td>Y</td>
<td>S</td>
<td>NA</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td>Oscillatoria agardhii CYA 29</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td><strong>Subsection V</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fischerella sp. PCC 7521</td>
<td>Y</td>
<td>S</td>
<td>none</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td><strong>Subsection IV</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anabaena sp. PCC 7108</td>
<td>Y</td>
<td>D</td>
<td>xisC (800 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Anabaena sp. PCC 7122</td>
<td>Y</td>
<td>D</td>
<td>xisC (500 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Anabaena variabilis</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Happe et al. (2000)</td>
</tr>
<tr>
<td>Cylindrospermum sp. PCC 7604</td>
<td>Y</td>
<td>D</td>
<td>xisC (700 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Cylindrospermum sp. PCC 7417</td>
<td>Y</td>
<td>D</td>
<td>xisC (700 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Nodularia sp. PCC 73104</td>
<td>Y</td>
<td>D</td>
<td>xisC (700 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Nostoc sp. PCC 6411</td>
<td>Y</td>
<td>D</td>
<td>xisC (700 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7121</td>
<td>Y</td>
<td>D</td>
<td>xisC (650 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7118</td>
<td>Y</td>
<td>D</td>
<td>xisC (750 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Nostoc punctiforme</td>
<td>N</td>
<td>D</td>
<td>NA</td>
<td>Meeks et al. (2001)</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7906</td>
<td>Y</td>
<td>D</td>
<td>xisC (600 bp)</td>
<td>This study</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7120</td>
<td>Y</td>
<td>D</td>
<td>xisC (600 bp)</td>
<td>Kaneko et al. (2001)</td>
</tr>
<tr>
<td>Nostoc sp. HCC 1048</td>
<td>Y</td>
<td>S</td>
<td>none</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td>Nostoc sp. HCC 1061</td>
<td>Y</td>
<td>S</td>
<td>none</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td>Nostoc sp. HCC 1075</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7107</td>
<td>Y</td>
<td>S</td>
<td>none</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td>Nostoc sp. PCC 6705</td>
<td>Y</td>
<td>S</td>
<td>none</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td>Nostoc sp. PCC 6314</td>
<td>Y</td>
<td>S</td>
<td>none</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td>Nostoc sp. PCC 6314/1</td>
<td>Y</td>
<td>S</td>
<td>none</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td>Nostoc muscorum CCAP 1453/12</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td>Nostoc sp. CYA 190</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td>Nostoc sp. CYA 238</td>
<td>Y</td>
<td>S</td>
<td>none</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td>Nostoc sp. CYA 295</td>
<td>Y</td>
<td>S</td>
<td>none</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td>Nostoc sp. CYA 306</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td>Nostoc edaphicum ACOI 97</td>
<td>Y</td>
<td>S</td>
<td>none</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td>Nostoc microscopicum ACOI 578</td>
<td>N</td>
<td>S</td>
<td>NA</td>
<td>Tamagnini et al. (2000)</td>
</tr>
<tr>
<td>Nostoc ellipsosporum ACOI 610</td>
<td>Y</td>
<td>S</td>
<td>none</td>
<td>Tamagnini et al. (2000)</td>
</tr>
</tbody>
</table>

were searched included NR, EST, WGS and individual genomes. All elements were searched against completely sequenced microbial genomes, including eubacteria and archaea.

Phylogenetic analyses of xisA and xisC nucleic acid sequences were accomplished with PAUP* 4.0b10 using parsimony, distance and maximum-likelihood (ML) criteria. The xisC and xisA from Nostoc sp. strain PCC 7210 and the xisC tree was rooted with xisC from Nostoc sp. strain PCC 7210. To construct the data matrices, inferred amino acid sequences were aligned using CLUSTAL W (Thompson et al., 1994) with gaps inserted for optimal alignment. The amino acid alignments were manually adjusted using MacClade 4.0. Nucleotide alignments were generated using Codon Align 2.0 (Hall, 2004), which constructs the nucleotide alignment based on the amino acid alignment, with gaps inserted between codons rather than within them.

Prior to performing ML analyses, Modeltest 3.7 was used to determine the evolutionary model that best fitted the xisA and xisC data (Posada & Crandall, 1998). Modeltest uses the hierarchical likelihood ratio test (hLRTs) and the Akaike information criterion (AIC) for selecting the best fit model. For xisA, Modeltest selected TrN+G (−lnL=7176.2539) for the hLRTs and GTR+G (−lnL=7145.8521) for AIC. For xisC, Modeltest selected HKY+G (−lnL=4562.8667) for the hLRTs and GTR+G (−lnL=4544.7305) for the AIC.

ML analyses were conducted using the heuristic search option, with the likelihood settings corresponding to each of the models selected by Modeltest. The parameters were as follows: no molecular clock was enforced, starting branch lengths were determined via the Rogers–Swofford approximation method, trees with likelihoods that were 5% or further from the target score were rejected, branch-length optimization equalled one-dimensional Newton Raphson with pass where the limit=20 and Δ=1e−06; starting trees were obtained via stepwise addition, sequence addition was random, one tree was held at each step during stepwise addition, the branch swapping algorithm tree bisection reconnection (TBR) was used, steepest descent option was not in effect, maxtree was set to 50 000, branches were collapsed if the branch lengths were equal to or less than 1e−08, the multrees option was in effect, and topological constraints were not enforced. Bootstrap values were calculated for 1000 replicates to evaluate branch support using parsimony criteria (Bremer, 1994; Felsenstein, 1985; Huelsenbeck et al., 1995).
RESULTS

Structural variation of the nifD element

To date, the nifD element has been completely sequenced from six cyanobacterial strains (Table 3), and there is considerable variation in size (ranging from 4 to 24 kb) and sequence composition among them (Henson et al., 2005) (Fig. 1). With the exception of the nifD elements from Nostoc sp. strain PCC 7120 and Anabaena variabilis, which are very similar to each other, the elements from the other strains were unique. There were conserved regions within all nifD elements (Fig. 1). The 5′ and 3′ flanking regions were conserved (Fig. 1) and each element was flanked by 11 bp direct repeats (CGGAGTAATCC).

However, in Anabaena sp. strain ATCC 33047 (Henson et al., 2005) and Nostoc sp. strain PCC 7416, the direct repeat nearest xisA differed by one nucleotide each, CGGAGTAATC in Anabaena sp. strain ATCC 33047 and CGGAGTAATCA in Nostoc sp. strain PCC 7416 (the variable nucleotide is in bold face and underscored). There were regions shared among individual elements that were not universally present (Fig. 1). For example, the hypothetical protein alr1449, originally designated ORF2 (Lammers et al., 1990; Henson et al., 2005) was conserved in all nifD elements, except in Nodularia spumigena CCY 9414 (Fig. 1). Others have suggested that alr1449 may be expressed in Nostoc sp. strain PCC 7120 (Rice et al., 1982). While some members of the serine family of site-specific recombinases require additional proteins for recombination, members of the tyrosine family, including xisA and xisC, do not (Smith & Thorpe, 2002; Nunes-Duby et al., 1998). This information, combined with its absence in Nodularia spumigena CCY 9414, suggests that alr1449 is not required for excision. The Nostoc sp. strain PCC 7120 and A. variabilis nifD elements were ~70% identical over their entire length, with the only difference being the absence of alr1443, asr1451, asr1452 and asr1453 in A. variabilis (Fig. 1). The element in Nostoc punctiforme contained a portion of all1445, which it shared with Nostoc sp. strain PCC 7120 and A. variabilis (Fig. 1). The nifD elements from Cylindrospermum sp. strain PCC 7417 and Nostoc punctiforme both contained the tRNA-Asn and ctiA genes. The tRNA-Asn gene encodes the tRNA for asparagine. The function of the ctiA gene remains unknown, but it is similar to the family of membrane-bound stomatin/prohibitin genes, which are found throughout the eukaryotes and the prokaryotes. The tRNA-Asn and ctiA region was conserved in A. variabilis, Nostoc sp. strain PCC 7120, Nostoc punctiforme and Cylindrospermum sp. strain PCC 7417; although, it was only located within the nifD element of the latter two.

BLAST searches of the sequenced nifD elements confirmed the expected similarities to the conserved regions of other

Table 3. Sequences used in this study and their GenBank accession numbers

<table>
<thead>
<tr>
<th>Strain</th>
<th>nifD element</th>
<th>hupL element</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anabaena sp. ATCC 33047</td>
<td>AY299473</td>
<td>NA</td>
</tr>
<tr>
<td>Anabaena sp. PCC 7108</td>
<td>HQ284169</td>
<td>HQ284178</td>
</tr>
<tr>
<td>Anabaena sp. PCC 7122</td>
<td>NA</td>
<td>HQ284177</td>
</tr>
<tr>
<td>Anabaena variabilis</td>
<td>NC_007413*</td>
<td>NA</td>
</tr>
<tr>
<td>Cylindrospermum sp. PCC 7417</td>
<td>HQ284186</td>
<td>HQ284186</td>
</tr>
<tr>
<td>Cylindrospermum sp. PCC 7604</td>
<td>HQ284176</td>
<td>HQ284184</td>
</tr>
<tr>
<td>Nodularia sp. PCC 73104</td>
<td>NA</td>
<td>HQ284182</td>
</tr>
<tr>
<td>Nodularia spumigena CCY 9414</td>
<td>NA</td>
<td>HQ284182</td>
</tr>
<tr>
<td>Nostoc punctiforme</td>
<td>NC_010628*</td>
<td>NA</td>
</tr>
<tr>
<td>Nostoc sp. PCC 6314</td>
<td>HQ284167</td>
<td>NA</td>
</tr>
<tr>
<td>Nostoc sp. PCC 6411</td>
<td>NA</td>
<td>HQ284179</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7118</td>
<td>HQ284170</td>
<td>HQ284180</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7120</td>
<td>NC_003272*</td>
<td>NC_003272*</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7121</td>
<td>HQ284174</td>
<td>HQ284181</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7416</td>
<td>HQ284168</td>
<td>NA</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7524</td>
<td>HQ284172</td>
<td>NA</td>
</tr>
<tr>
<td>Nostoc sp. PCC 7906</td>
<td>HQ284175</td>
<td>HQ284184</td>
</tr>
<tr>
<td>Scytonema sp. PCC 7110</td>
<td>HQ284173</td>
<td>NA</td>
</tr>
<tr>
<td>Tolypothrix PCC 7415</td>
<td>HQ284171</td>
<td>NA</td>
</tr>
</tbody>
</table>

*Genome sequences retrieved from GenBank and not derived from this study.
*nifD* elements (Fig. 1). They also revealed that particular regions of these elements were similar to a variety of other sequences not found in other *nifD* elements. For instance the hypothetical gene *alr1444* (present in the *nifD* elements of *Nostoc* sp. strain PCC 7120 and *A. variabilis*) was similar to a number of cyanobacterial genes including *pebAB* (involved in the synthesis of the photosynthetic pigment phycoerythrobilin (Alvey et al., 2003), *sll1232* in *Synechocystis* sp. strain PCC 6803, and *glr0577* and *gll1896* in *Gloeobacter* sp. strain PCC 7421. The *asl1446* gene, present in the *nifD* elements of *Nostoc* sp. strain PCC 7120 and *A. variabilis*, was similar to putative proteins within a number of cyanobacteria and, in some cases, to multiple genes within individual genomes. In addition, a region of the *nifD* element in *A. variabilis* was similar to *rbpG*, which encodes an RNA-binding protein that is located in multiple places within the *A. variabilis* genome (Maruyama et al., 1999).

In addition to the *tRNA-Asn* and *ctiA* genes mentioned earlier, the 24 kb *nifD* element in *Nostoc punctiforme* contained regions that were similar to eubacterial kinases, the *Synechocystis* sp. strain PCC 6803 *slr1963* gene that is believed to encode solanesyl diphosphate synthase involved in the formation of isoprenoid quinones (Okada et al., 1997), and the *Nostoc* sp. strain PCC 7120 *alr1077* gene, which codes for a probable carboxymethylenbutenolidase (Sazuka, 2003).

The *Nodularia spumigena* CCY 9414 *nifD* element contained a 200 bp region which was similar to the *Nostoc* sp. strain PCC 7524 *NspV* endonuclease. *BLAST* searches of the *nifD* element of *Anabaena* sp. strain ATCC 33047 revealed a 125 bp region of this element, composed mostly of repeating CATCTT units, which was similar to several viruses (Rabbit fibroma virus and Rat cytomegalovirus) (Henson et al., 2005). Other than the conserved regions found in the other sequenced *nifD* elements, the 9 kb *nifD* element from *Cylindrospermum* sp. strain PCC 7417 had little similarity to any other known sequence.

**Fig. 1.** A diagram comparing the similarity between *nifD* elements. All ORFs of the *Nostoc* sp. strain PCC 7120 *nifD* element (top) are represented by labelled boxes with different graphical patterns. The direct repeats are represented by vertical arrows and the horizontal arrows indicate the direction of transcription for *nifD* and *xisA*. Regions of the *A. variabilis*, *Anabaena* sp. strain ATCC 33047, *Nostoc punctiforme*, *Cylindrospermum* sp. strain PCC 7417 and *Nodularia spumigena* CCY 9414 *nifD* elements that are similar to *Nostoc* sp. strain PCC 7120 genes are represented by boxes with identical graphical patterns and intergenic regions are represented by horizontal bars.

**Structural variation of the *hupL* element**

To date, the *hupL* element from three cyanobacterial strains has been sequenced and there was variation in size (ranging from 7 to 9.4 kb) and composition among them. Most of each element was unique; but there were also conserved regions (Fig. 2). The 5′ and 3′ flanking regions, including *xisC*, were conserved. The *hupL* element was flanked by 16 bp direct repeats, CCATATACTGCTGTG. However, in the *hupL* elements of *Cylindrospermum* sp. strain PCC 7417 and *Nodularia spumigena* CCY 9414, the direct repeat nearest *xisC* differed by one nucleotide, CCATATACTGCTGTG (variable nucleotide in bold and underscored). The region encompassing *asr0680*, *alr0681* and *asr0682* was conserved in all *hupL* elements, however the
arrangement of these hypothetical genes was different in *Cylindrospermum* sp. strain PCC 7417 compared with the other two strains. In *Nostoc* sp. strain PCC 7120 and *Nodularia spumigena* CCY 9414, they were encoded as *asr0680*, *alr0681* and *asr0682* (5' to 3'), whereas in *Cylindrospermum* sp. strain PCC 7417 they were encoded in the order: *asr0682*, *asr0680* and *alr0681* (Fig. 2). Although it cannot be resolved with absolute certainty, we hypothesize that *asr0682* (as it appears in *Cylindrospermum* sp. strain PCC 7417) is the result of a translocation. This is supported by comparative sequence analysis of the three sequenced *hupL* elements. The intergenic region between *asr0681* and *asr0682* (present in *Nodularia spumigena* CCY 9414 and *Nostoc* sp. strain PCC 7120) was absent in *Cylindrospermum* sp. strain PCC 7417. In *Cylindrospermum* sp. strain PCC 7417, the end of *asr0682* was located exactly 92 bp upstream of the start of *asr0680*. This 92 bp region corresponded to the region just upstream of *asr0680* in *Nostoc* sp. strain PCC 7120 and *Nodularia spumigena* CCY 9414. These data supported the hypothesis that the location of *asr0682* in *Cylindrospermum* sp. strain PCC 7417 is the result of a translocation. In addition, *Nodularia spumigena* CCY 9414 had a 1.5 kb segment of DNA that was also similar (85%) to the *all0684* and *alr0686* of *Nostoc* sp. strain PCC 7120 (Fig. 2).

**Phylogenetic analysis of xisA and xisC**

The *xisA* aligned data matrix was 1302 characters in length (see Supplementary Fig. S1 in IJSEM Online) and ML analyses using the TrN + G (−lnL=3772.39389) and GTR + G models (−lnL=3747.13269) produced two trees each, all four of which had similar topologies. Analysis of *xisA* produced results that were consistent with other cyanobacterial phylogenies (Henson et al., 2004). A strict consensus of the two trees derived from the TrN + G method is presented in Fig. 3. This tree consists of two major clades, with *Cylindrospermum* sp. strains PCC 7417 and PCC 7604 at the base of the tree. The first clade is composed of all of the *Anabaena* and most of the *Nostoc* strains, interspersed within the clade. The remaining *Nostoc* strains are placed in a clade along with the *Scytonema* and *Nodularia* strains.

The *xisC* aligned data matrix was 1551 characters (see Supplementary Fig. S2) in length, and ML analyses using the HKY + G (−lnL=2919.18578) and GTR + G models (−lnL=2896.23888) resulted in 61 and 26 trees, respectively, all of which had similar topologies. A strict consensus of the 61 trees generated by the HKY + G method is presented in Fig. 4. Within this tree, the *Nostoc* and *Anabaena* strains occurred in an unresolved polytomy at the base of the tree, sister to a clade containing the *Cylindrospermum* and *Nodularia* strains. Although analysis of *xisC* lacked resolution, the relationships among the strains were congruent with other published cyanobacterial phylogenies (Henson et al., 2004).

---

**Fig. 2.** A diagram showing the similarity between the *hupL* elements of *Nostoc* sp. strain PCC 7120, *Nodularia spumigena* CCY 9414 and *Cylindrospermum* sp. strain PCC 7417. All genes are labelled and represented by boxes that have different graphical patterns. The direct repeats are represented by vertical arrows and the horizontal arrows indicate the direction of transcription for *hupL* and *xisC*. Regions of the *Nodularia spumigena* CCY 9414 and *Cylindrospermum* sp. strain PCC 7417 elements that are similar to *Nostoc* sp. strain PCC 7120 are represented by boxes with identical graphical patterns and intergenic regions are represented by horizontal bars.
DISCUSSION

Heterocyst differentiation in cyanobacteria is initiated by nitrogen deprivation. The transcription factor NtcA can gauge levels of environmental nitrogen levels (Muro-Pastor et al., 2001; Tanigawa et al., 2002; Herrero et al., 2004) and is believed to be the key regulator in initiating heterocyst differentiation and nitrogen fixation (Chastain et al., 1990; Jiang et al., 2000; Ramasubramanian et al., 1994, 1996; Herrero et al., 2004). Factors that induce the expression of \textit{xisC} are unknown, however, the \textit{xisA} promoter has binding sites for NtcA and FurA, suggesting that these transcription factors are involved in expression of \textit{xisA} (Jiang et al., 2000; Ramasubramanian et al., 1994, 1996; Herrero et al., 2004; López-Gomollón et al., 2007). Factors that induce the expression of \textit{xisC} are unknown, however, the \textit{xisA} promoter has binding sites for NtcA and FurA, suggesting that these transcription factors are involved in expression of \textit{xisA} (Jiang et al., 2000; Ramasubramanian et al., 1994, 1996; Herrero et al., 2004; López-Gomollón et al., 2007). Understanding the evolutionary origin of the elements and their genes that encode self-excision may also provide insight into the origin and evolution of heterocyst formation.

There is considerable variation in size and composition within the representative \textit{nifD} and \textit{hupL} elements. The \textit{nifD} element ranges from 4 to 24 kb and the \textit{hupL} element ranges from 7 to 9.4 kb. Neither element has much sequence similarity with other published sequences. However, there are conserved regions within each representative \textit{nifD} and \textit{hupL} element. Both element types have conserved 5' and 3' flanking regions, each element type encodes its own recombinase and, with the exception of \textit{Nodularia spumigena} CCY 9414, each element type has a conserved central region. Phylogenetic analysis of \textit{xisA} and \textit{xisC} sequences resulted in phylogenies that were congruent with published heterocystous cyanobacterial phylogenies (Henson et al., 2004). This suggests that subsequent to the elements' initial insertion into the cyanobacterial lineage, they have undergone vertical descent and have not been transferred laterally within the lineage.

The evolutionary origin of the \textit{nifD} and \textit{hupL} elements remains unclear. The data suggest that they are the remnants of viruses or bacteriophages. This is supported by several lines of evidence. XisA and XisC belong to the tyrosine family of recombinases, which are mostly found in viruses (Nunes-Duby et al., 1998; Carrasco et al., 2005). In addition, the XisA and XisC sequences are more similar to each other than any other published sequence and both have the tyrosine to histidine mutation in the active site, suggesting that XisA and XisC, as well as the \textit{nifD} and \textit{hupL} elements, share a common evolutionary origin.

Several structural aspects of the \textit{nifD} and \textit{hupL} elements suggest that they have a viral ancestry and may be bacteriophages that have lost the ability to self-rePLICATE. The location of the XisA and XisC recombinases near the 5' end of the \textit{nifD} and \textit{hupL} elements and the fact that the

Fig. 3. A phylogenetic tree of \textit{xisA} sequences created using ML analysis. Bootstrap values are listed above the branches.
elements are inserted in exactly the same location within the nifD and hupL genes in all examined strains suggest a viral origin (Casjens, 2003). However, the only segment of any sequenced nifD or hupL element (other than xisA and xisC) that is similar to a virus is the 125 bp region of the Anabaena sp. strain ATCC 33047 nifD element, which is similar to the Rabbit fibroma virus and Rat cytomegalovirus (Henson et al., 2005). This could suggest that the nifD or hupL elements are not of viral origin; however, these elements may represent defective prophages that are in a state of mutational decay (Casjens, 2003; Canchaya et al., 2003). Defective prophages in a state of mutational decay are characterized by having increased mutational rates in select regions as well as having large-scale deletions (Casjens, 2003; Canchaya et al., 2003). In defective prophages, it is possible for certain genes to be retained through selection, while the remainder of the prophage accumulates mutations. This could explain why there is so much structural and compositional variation within the nifD and hupL elements, yet certain regions remain conserved. It is also possible for a highly mutated prophage to have diverged so much that no characteristic genes from the virus, other than the recombine, remain (Casjens, 2003).

If the nifD and hupL elements are defective prophages, they exist as part of the flexible gene pool in cyanobacteria. The flexible gene pool is best described as the DNA within a bacterial genome that arose through horizontal transfer (Hacker & Carniel, 2001). The flexible gene pool consists of genetic entities such as genomic islands, prophages, integrons, transposons, genomic islets and plasmids (Hacker & Carniel, 2001). Genetic information within the flexible gene pool may provide an ecological or pathogenic advantage; however, it could also provide no selective advantage and exists as selfish DNA (Lilley et al., 2000).

Although the nifD and hupL elements may be descended from a common ancestor, their insertion into the nifD and hupL genes probably occurred as two separate events. If the nifD and hupL elements resulted from a single insertion event (i.e., a single origin) into the cyanobacteria and subsequently underwent duplication and divergence, we would expect more sequence conservation between the two than is currently present. Other than xisA and xisC, which have similar nucleotide and amino acid sequences, the nifD elements share no sequence similarity to the hupL elements. In addition to their independent insertions, the nifD and

---

**Fig. 4.** A phylogenetic tree of xisC sequences created using ML analysis. Bootstrap values are listed above the branches.
hupL elements have probably been under different selection pressures, which might explain the haphazard occurrence of the elements within the cyanobacteria. Initially both elements were thought to only be in Subsection IV, which is the case for the nifD element. However, the hupL element (or at least a portion of xisC) has been found within a Subsection V Fischerella strain and a Subsection III Leptolyngbya strain (Tamagnini et al., 2000), suggesting that the hupL element may have originated within the lineage prior to the diversification of the filamentous cyanobacteria (Subsections III, IV and V).

Individual regions within each element type may be under different selection pressures and have experienced differential rates of evolution. This would explain the extensive structural and compositional variation within the nifD and hupL elements that is coupled with the presence of conserved regions. If only specific regions of the nifD and hupL elements are under selective pressure, then the remaining portions of each element would be free to accumulate mutations (e.g. insertions, deletions, and translocations) without inhibiting excision or being detrimental to the organism.

For DNA elements to be maintained in the cyanobacterial genomes, in theory there should be some sort of selective pressure to retain them. Although the genomes, in theory, there should be some sort of selective pressure to maintain DNA elements in the cyanobacterial genome rearrangements. Detrimental to the organism.

For DNA elements to be maintained in the cyanobacterial genomes, in theory there should be some sort of selective pressure to retain them. Although the genomes, in theory, there should be some sort of selective pressure to maintain DNA elements in the cyanobacterial genome rearrangements. Detrimental to the organism.

The variation in size, structure, and composition of nifD and hupL elements indicates a complex pattern of insertions, deletions, translocation and sequence divergence over the course of evolution. This complex evolutionary pattern suggests that, aside from the conserved regions, these elements are a repository of genetic information where segments of DNA can undergo mutations with little to no effect on the organism or element excision. The nifD and hupL elements represent an interesting aspect of the regulation of nitrogen fixation, heterocyst differentiation and the molecular evolution of genome structure within the heterocystous cyanobacteria. In addition, they may represent an excellent system for the study of defective prophages that are in advanced stages of mutational decay.

REFERENCES


