Protein signatures (molecular synapomorphies) that are distinctive characteristics of the major cyanobacterial clades

Radhey S. Gupta

Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario L8N 3Z5, Canada

A combination of phylogenomic and signature sequence-based (or phenetic) approaches was used to understand the evolutionary relationships among cyanobacteria. Phylogenetic trees were constructed for 34 cyanobacteria whose genomes have been sequenced, based on concatenated sequences for 45 conserved proteins and also the 16S rRNA gene. In parallel, sequence alignments of various proteins were examined to identify conserved indels (i.e. molecular signatures or synapomorphies) that are specific for either all cyanobacteria or their various clades in the phylogenetic trees. Of the >40 molecular signatures described in this work, 15 are specific for all cyanobacteria. The other cyanobacterial clades that can now be identified and circumscribed in molecular terms by using these signatures include a deep-branching clade (clade A, corresponding to the subclass Gloeobacterophycidae), consisting of *Gloeobacter violaceus* and two diazotrophic *Synechococcus* strains (JA-3-3Ab and JA2-3-Ba) (15 aa insert in EF-G); a clade comprising all other cyanobacteria except those from clade A [18 aa insert in DNA polymerase I (Pol I), 2 aa insert in the DnaX protein, 4 aa insert in TrpRS and 4–5 aa insert in tryptophan synthase beta subunit]; a clade (clade C, corresponding to the subclass *Synechococcophycidae*) of various marine unicellular *Synechococcus* and *Prochlorococcus* cyanobacteria (12 aa insert in Pol I, 3 aa insert in RpoB, 2 aa insert in KgsA, 6 aa insert in TyrRS, 2 aa insert in rRNA-mG1 transferase and 1 aa deletion in the RpoC protein); a clade of the low-B/A ecotype *Prochlorococcus* strains (5 aa deletion in LeuRS and 1 aa insert in the Ffh protein); a clade consisting of the *Nostocales* species/strains (subclass Nostocophycidae; 4 aa insert in the PetA protein and 5 aa insert in the ribosomal protein S3); a clade of the order *Chroococcales* (1 aa insert in RecA); a clade comprising the orders *Nostocales*, *Oscillatoriales* and *Chroococcales* [19 aa insert in DnaE, 13 aa insert in GDP–mannose pyrophosphorylase and 22–27 aa insert in NADP(H)–quinone oxidoreductase subunit D]. Two additional conserved indels in the translation-initiation factor IF-2 and riboflavin synthase alpha subunit suggest an intermediate placement of the *Oscillatoriales* in between the orders *Nostocales* and *Chroococcales*. The unique presence of these molecular signatures in all available sequences from the indicated groups of cyanobacteria, but not in any other cyanobacteria (or bacteria), indicates that these synapomorphies provide novel and potentially useful means for circumscription of several important taxonomic clades of cyanobacteria in more definitive terms. The species-distribution patterns of these synapomorphies also indicate that the plant/plastid homologues are not derived from the clade A or C cyanobacteria.

INTRODUCTION

Cyanobacteria form one of the main phyla within *Bacteria* and they are the sole prokaryotic group capable of carrying out oxygenic photosynthesis (Kondratieva *et al.*, 1992; Castenholz, 2001). They exhibit enormous diversity in terms of their morphology (cell size, shape and arrangement, e.g. filamentous, unicellular or colonial), physiology (e.g. ability to fix nitrogen, use of sulfide as an electron donor) and other characteristics (e.g. motility, thermophily) (Rippka *et al.*, 1979; Anagnostidis & Komarek, 1985; Wilmotte & Golubic, 1991; Kondratieva *et al.*, 1992; Castenholz, 2001; Sánchez-Baracaldo *et al.*, 2005). However, most of these characteristics in the past have
not shown good correlation with cyanobacterial phylogeny based on the 16S rRNA gene or other gene/protein sequences (Honda et al., 1999; Turner et al., 1999; Ishida et al., 2001; Robertson et al., 2001; Wilmotte & Herdman, 2001; Gugger & Hoffmann, 2004; Sánchez-Baracaldo et al., 2005; Swingley et al., 2008a). Thus, it remains to be determined whether some of them will provide reliable markers for understanding cyanobacterial taxonomy or evolution. In 16S rRNA gene-based trees, which provide the primary means for taxonomic and evolutionary studies (Olsen & Woese, 1993; Honda et al., 1999; Turner et al., 1999; Wilmotte & Herdman, 2001), cyanobacteria group into 14 clusters (Wilmotte & Herdman, 2001). However, the taxonomic significances of these clusters or how they are related evolutionarily is not clear (Honda et al., 1999; Turner et al., 1999; Wilmotte & Herdman, 2001; Hoffmann, 2005). Apart from the 16S rRNA gene, sequence information for cyanobacteria for other genes/proteins was, until recently, quite limited; however, phylogenetic analyses based on them generally indicated results very similar to those seen with the 16S rRNA gene (Giovannoni et al., 1988; Delwiche et al., 1995; Honda et al., 1999; Henson et al., 2002; Seo & Yokota, 2003; Zhaxybayeva et al., 2006; Shi & Falkowski, 2008). Many of the problems that are currently encountered in understanding cyanobacterial phylogeny can be attributed to their outdated taxonomy and nomenclature, which have been governed in the past by both the Botanical and Bacteriological Codes (Rippka et al., 1979; Anagnostidis & Komarek, 1985; Castenholz, 2001; Oren, 2004; Garrity et al., 2005). Thus, despite the fact that cyanobacteria constitute a large phylum containing >4000 isolates (Maidak et al., 2001), only a small number of species and higher taxonomic groups within this phylum have validly published names under the Bacteriological Code (Oren, 2004; Garrity et al., 2005; Hoffmann, 2005). At the 16th Symposium of the International Association for Cyanophyte Research held in 2004, which considered some of these problems, a number of recommendations for improvement of cyanobacterial nomenclature and for their integration under the Bacteriological Code were proposed (Oren, 2004; Hoffmann, 2005; Oren & Tindall, 2005). At this meeting, a new proposal for the classification of cyanobacteria into four subclasses, viz. Gloeobacteriophyceae, Synechococccophycideae, Oscillariariophyceae and Nostocophyceae, was also made (Hoffmann et al., 2005).

The availability of genome sequences in recent years has provided new opportunities for understanding cyanobacterial taxonomy and evolution. Based upon genomic sequences, an important and frequently used approach is to assemble phylogenies based upon combined sequences for a large number of proteins (Eisen, 1998; Ciccarelli et al., 2006). Phylogenies based upon a large number of characters derived from multiple conserved (or slow-evolving) genes/proteins are better able to resolve deeper-branching evolutionary relationships than those based on any single gene or protein (Hansmann & Martin, 2000; Brown et al., 2001; Rokas et al., 2003; Delsuc et al., 2005; Ciccarelli et al., 2006). Recently, a number of studies have reported phylogenomic analysis for limited numbers of cyanobacteria (between 11 and 24) based upon concatenated sequences for different large sets of proteins present in various cyanobacteria (Giovannoni et al., 1988; Honda et al., 1999; Wilmotte & Herdman, 2001; Seo & Yokota, 2003; Sánchez-Baracaldo et al., 2005; Shi & Falkowski, 2008; Swingley et al., 2008a). These studies have been very useful in confirming the existence of certain important clades within cyanobacteria and also in clarifying their relative branching positions. However, based upon these studies, one cannot circumscribe various cyanobacterial clades in definitive terms, which is a prerequisite for developing reliable or stable taxonomy (Oren & Stackebrandt, 2002; Hoffmann et al., 2005). This approach is also limited to species whose genomes have been sequenced.

Hence, it is important to identify other reliable molecular markers that are consistent with the results of 16S rRNA gene-based trees as well as phylogenomic and other approaches, but which also enable circumscription of the major cyanobacterial clades in more definitive (molecular) terms. One approach that has been very useful in this regard involves identifying molecular synapomorphies consisting of conserved inserts or deletions (i.e. indels) in widely distributed proteins that are distinctive characteristics of various main clades within a given phylum (Rivera & Lake, 1992; Delwiche et al., 1995; Gupta, 1998, 2000; Griffiths & Gupta, 2007). This approach has been used successfully in our work to elucidate the branching order and inter-relationships within different main phyla of bacteria (Gupta et al., 1999; Griffiths & Gupta, 2001; Gupta, 2001, 2003), as well as a number of main groups (or phyla) of bacteria including Alphaproteobacteria (Gupta & Mok, 2007), Chlamydiae (Gupta & Griffiths, 2006) and the Bacteroidetes–Chlorobi group (Gupta, 2004). By using this approach, we have previously described 14 conserved indels in 10 widely distributed proteins that were distinctive characteristics of various available cyanobacteria, but not found in any other bacteria (Gupta et al., 2003). The shared presence of several of these indels by the plastid/plant homologues (Gupta et al., 2003) also provided evidence for their cyanobacterial ancestry (Margulis, 1970; Gray & Doolittle, 1982; Morden et al., 1992; Whatley, 1993; Palmer & Delwiche, 1998). In this work, we describe 28 novel conserved indels in many important and widely distributed proteins that are distinctive characteristics of various main clades of cyanobacteria that can be identified by means of phylogenetic analyses. Because of their specific presence in these clades, these conserved indels or synapomorphies also provide new means for potential circumscription of a number of important taxonomic groups of cyanobacteria in molecular terms and for understanding cyanobacterial evolution.
METHODS

Phylogenetic analyses. Phylogenetic analyses were carried out on a set of proteins involved in important housekeeping functions that are present in most organisms (Harris et al., 2003). BLAST searches were conducted on each of these proteins to determine whether their homologues were present in all 34 cyanobacteria listed in Table 1 and the two outgroup species (Bacillus subtilis and Staphylococcus aureus) used in this work. Except for Crocosphaera watsonii WH8501, the genomes of all other cyanobacteria are now available (Table 1). For the 45 proteins for which information is presented in Supplementary Table S1 (available in IJSEM Online), orthologues were available from all 34 cyanobacteria and the two outgroup species. The multiple sequence alignments for these proteins were created by using the CLUSTAL_X 1.83 program (Jeanmougin et al., 1998) and they were concatenated into a single large file. This unedited sequence alignment was imported into the Gblocks 0.91b program (Castresana, 2000) to remove poorly aligned regions. This program was used with default settings except the allowed gap position parameter was changed to 0.5. The resulting final alignment of 16,822 amino acid sites was used for phylogenetic analyses. A neighbour-joining (NJ) tree based on 1000 bootstrap replicates was constructed by the Kimura model (Kimura, 1983) using the TREECON 1.3b program (Van de Peer & De Wachter, 1994). The maximum

Table 1. Cyanobacterial genomes studied in this work

Abbreviations: DOE JGI, Department of Energy Joint Genome Institute; GBM, Gordon and Betty Moore; TIGR, The Institute of Genome Research.
The genome of Crocosphaera watsonii WH8501 was not sequenced fully.

<table>
<thead>
<tr>
<th>Species/strain</th>
<th>Genome size (Mb)</th>
<th>DNA G+C content (mol%)</th>
<th>No. proteins*</th>
<th>GenBank accession no.</th>
<th>Centre/reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acaryochloris marina MBIC11017</td>
<td>8.36</td>
<td>47.0</td>
<td>6254</td>
<td>NC_009925.1</td>
<td>Swingley et al. (2008b)</td>
</tr>
<tr>
<td>Anabaena variabilis ATCC 29413</td>
<td>7.07</td>
<td>41.4</td>
<td>5043</td>
<td>NC_007413.1</td>
<td>DOE JGI</td>
</tr>
<tr>
<td>Gloeobacter violaceus PCC7421</td>
<td>4.66</td>
<td>62.0</td>
<td>4430</td>
<td>NC_005125.1</td>
<td>Nakamura et al. (2003)</td>
</tr>
<tr>
<td>Nostoc sp. PCC7120</td>
<td>7.21</td>
<td>41.3</td>
<td>5366</td>
<td>NC_003272.1</td>
<td>Kaneko et al. (2001)</td>
</tr>
<tr>
<td>Nostoc punctiforme PCC73102</td>
<td>8.20</td>
<td>41.4</td>
<td>6087</td>
<td>NC_010628</td>
<td>DOE JGI</td>
</tr>
<tr>
<td>Prochlorococcus marinus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AS9601</td>
<td>1.70</td>
<td>31.3</td>
<td>1921</td>
<td>NC_008816.1</td>
<td>J. Craig Venter Institute</td>
</tr>
<tr>
<td>MIT9211</td>
<td>1.70</td>
<td>39.7</td>
<td>1855</td>
<td>NC_009976.1</td>
<td>Rocap et al. (2003)</td>
</tr>
<tr>
<td>MIT9215</td>
<td>1.70</td>
<td>31.0</td>
<td>1983</td>
<td>NC_009840.1</td>
<td>DOE JGI</td>
</tr>
<tr>
<td>MIT9301</td>
<td>1.60</td>
<td>31.3</td>
<td>1907</td>
<td>NC_009091.1</td>
<td>GBM Foundation</td>
</tr>
<tr>
<td>MIT9303</td>
<td>2.70</td>
<td>50.0</td>
<td>2997</td>
<td>NC_008820.1</td>
<td>J. Craig Venter Institute</td>
</tr>
<tr>
<td>MIT9312</td>
<td>1.71</td>
<td>31.2</td>
<td>1810</td>
<td>NC_007577.1</td>
<td>DOE JGI</td>
</tr>
<tr>
<td>MIT9313</td>
<td>2.41</td>
<td>50.7</td>
<td>2269</td>
<td>NC_005071.1</td>
<td>Rocap et al. (2003)</td>
</tr>
<tr>
<td>MIT9515</td>
<td>1.70</td>
<td>30.8</td>
<td>1906</td>
<td>NC_008817.1</td>
<td>J. Craig Venter Institute</td>
</tr>
<tr>
<td>NATL1A</td>
<td>1.90</td>
<td>35.0</td>
<td>2193</td>
<td>NC_008819.1</td>
<td>J. Craig Venter Institute</td>
</tr>
<tr>
<td>NATL2A</td>
<td>1.80</td>
<td>35.1</td>
<td>2163</td>
<td>NC_007335.2</td>
<td>DOE JGI</td>
</tr>
<tr>
<td>Prochlorococcus marinus subsp. marinus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCMP1375</td>
<td>1.75</td>
<td>36.4</td>
<td>1883</td>
<td>NC_005042.1</td>
<td>Dufresne et al. (2003)</td>
</tr>
<tr>
<td>CCMP1986</td>
<td>1.70</td>
<td>30.8</td>
<td>1717</td>
<td>NC_005072.1</td>
<td>Rocap et al. (2003)</td>
</tr>
<tr>
<td>Synechococcus elongatus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCC6301</td>
<td>2.70</td>
<td>55.5</td>
<td>2527</td>
<td>NC_006576.1</td>
<td>Sugita et al. (2007)</td>
</tr>
<tr>
<td>PCC7942</td>
<td>2.75</td>
<td>55.4</td>
<td>2612</td>
<td>NC_007604.1</td>
<td>DOE JGI</td>
</tr>
<tr>
<td>Synechococcus sp.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC9311</td>
<td>2.61</td>
<td>52.4</td>
<td>2892</td>
<td>NC_008319.1</td>
<td>Palenik et al. (2006)</td>
</tr>
<tr>
<td>CC9605</td>
<td>2.51</td>
<td>59.2</td>
<td>2645</td>
<td>NC_007516.1</td>
<td>Dufresne et al. (2008)</td>
</tr>
<tr>
<td>CC9902</td>
<td>2.23</td>
<td>54.2</td>
<td>2307</td>
<td>NC_007513.1</td>
<td>Dufresne et al. (2008)</td>
</tr>
<tr>
<td>JA-2-3B’a(2-13)</td>
<td>3.05</td>
<td>58.5</td>
<td>2862</td>
<td>NC_007776.1</td>
<td>TIGR</td>
</tr>
<tr>
<td>JA-3-3Ab</td>
<td>2.93</td>
<td>60.2</td>
<td>2760</td>
<td>NC_007775.1</td>
<td>TIGR</td>
</tr>
<tr>
<td>RCC307</td>
<td>2.20</td>
<td>60.8</td>
<td>2535</td>
<td>NC_009482.1</td>
<td>Dufresne et al. (2008)</td>
</tr>
<tr>
<td>RS9917</td>
<td>2.58</td>
<td>64.8</td>
<td>2820</td>
<td>–</td>
<td>Dufresne et al. (2008)</td>
</tr>
<tr>
<td>WH5701</td>
<td>2.86</td>
<td>66.0</td>
<td>3129</td>
<td>–</td>
<td>Dufresne et al. (2008)</td>
</tr>
<tr>
<td>WH7803</td>
<td>2.40</td>
<td>60.2</td>
<td>2533</td>
<td>NC_009481.1</td>
<td>Dufresne et al. (2008)</td>
</tr>
<tr>
<td>WH7805</td>
<td>2.62</td>
<td>57.5</td>
<td>2934</td>
<td>–</td>
<td>Dufresne et al. (2008)</td>
</tr>
<tr>
<td>WH8102</td>
<td>2.43</td>
<td>59.4</td>
<td>2519</td>
<td>NC_005070.1</td>
<td>Palenik et al. (2003)</td>
</tr>
<tr>
<td>Synechocystis sp. PCC6803</td>
<td>3.95</td>
<td>47.4</td>
<td>3172</td>
<td>NC_000911.1</td>
<td>Kaneko et al. (1996)</td>
</tr>
<tr>
<td>Thermosynechococcus elongatus BP-1</td>
<td>2.59</td>
<td>53.9</td>
<td>2476</td>
<td>NC_004113.1</td>
<td>Nakamura et al. (2002)</td>
</tr>
<tr>
<td>Trichodesmium erythraeum IMS101</td>
<td>7.80</td>
<td>34.1</td>
<td>4451</td>
<td>NC_008312.1</td>
<td>DOE JGI</td>
</tr>
</tbody>
</table>

*Total number of proteins or open reading frames in the genome.
Likelihood (ML) analysis was carried out by using the WAG+F model with gamma distribution of evolutionary rates with four categories by using the TREE-PUZZLE program with 10,000 puzzling steps (Schmidt et al., 2002). The 16S rRNA gene sequences for the same cyanobacteria were obtained from their genomic sequences. A multiple sequence alignment of them was created by using the CLUSTAL_X 1.83 program. An NJ tree based on 1000 bootstrap replicates was constructed based on distances calculated using Kimura’s two-parameter model (Kimura, 1980) using the TREECON 1.3b program. This tree was also rooted by using B. subtilis and S. aureus sequences.

Identification of conserved indels that are specific for various cyanobacterial clades. To identify conserved indels that are specific for cyanobacteria, multiple sequence alignments of different proteins made in our earlier work (Gupta, 1998, 2004; Gupta & Griffiths, 2006; Gupta & Mok, 2007) were examined. BLAST searches were also performed on many other conserved proteins from the genome of Nostoc sp. PCC7120 to retrieve cyanobacterial homologues, as well as high-scoring homologues from other bacteria, to generate multiple sequence alignments. All of these sequence alignments were inspected visually to identify conserved indels that were restricted to cyanobacteria (Gupta et al., 2003). The indels that were not flanked by conserved regions were excluded from further consideration, as they do not provide reliable molecular markers (Gupta, 1998). The species distribution of all conserved indels was evaluated by carrying out detailed BLASTP searches on short sequence segments (between 60 and 150 aa depending upon the lengths of the indels) containing the indels and their flanking conserved regions. In cases where the indels were large, two separate BLASTP searches were performed, one with a sequence containing the indel and the other with a sequence lacking the indel, to determine the presence or absence of the indel in various species/strains. The sequence information for various conserved indels from representative cyanobacteria and some other bacteria was compiled into signature files. Due to space consideration, sequences for some closely related Prochlorococcus and Synechococcus strains are not shown in several of the alignment figures.

To infer whether a given indel is an insert or a deletion in cyanobacteria or their particular groups, the presence of this indel in other cyanobacteria, as well as other phyla of bacteria, was determined. If this indel was lacking in these groups, thereby indicating that the absence of indel is the ancestral character state, then the indel in question was interpreted as an insert in the given group of cyanobacteria. On the other hand, if all of these bacteria contained the particular sequence region that was absent in a given clade of cyanobacteria, then the indel was considered a deletion.

RESULTS

Phylogenomic phylogenetic analyses on cyanobacteria

Genome sequences for a large number of cyanobacteria have become available in recent years (Table 1). Phylogenetic analysis on these cyanobacteria was carried out based on a concatenated sequence alignment for 45 widely distributed proteins (see Methods). A rooted NJ tree based on this dataset is shown in Fig. 1 and an ML tree based on these sequences is provided as Supplementary Fig. S1, available in IJSEM Online. The branching patterns of cyanobacteria in these trees are very similar and these results are comparable to those observed in other recent studies based on different large datasets of protein sequences (Sánchez-Baracaldo et al., 2005; Shi & Falkowski, 2008; Swingley et al., 2008a). In both NJ and ML trees, a clade consisting of Gloeobacter violaceus and Synechococcus spp. (JA-3-3Ab and JA2-3-B’a) (referred to here as clade A) showed the deepest branching within the cyanobacteria. The deep branching of clade A species has also been observed in a number of earlier studies (Giovannoni et al., 1988; Honda et al., 1999; Wilmotte & Herdman, 2001; Seo & Yokota, 2003; Sánchez-Baracaldo et al., 2005; Shi & Falkowski, 2008; Swingley et al., 2008a).

According to a recent proposal for the classification of cyanobacteria (Hoffmann et al., 2005), the Gloeobacterales are placed into a separate subclass (Gloeobacterophycidae). Most other cyanobacteria in our dataset can be grouped into two major clades. One of these clades (designated clade B) comprises diverse cyanobacteria such as Thermosynechococcus, Synechocystis, Crocosphaera, Acaryochloris, Trichodesmium, Nostoc and Anabaena spp., and it includes both the subclasses Oscillariiophycidae and Nostocophycidae of cyanobacteria (Hoffmann et al., 2005). In contrast, the other main clade (clade C) is composed entirely of different strains/isolates of Prochlorococcus and Synechococcus. This latter clade corresponds to the subclass Synechococcomphycidae (Hoffmann et al., 2005), except that our analysis indicates that Acaryochloris is not part of this clade or subclass. Based on their genetic distances (Figs 1 and 2), the clade C isolates are genetically not as diverse as the taxa in clade B. Within clade C, two subclades can be distinguished. One of these subclades consisted entirely of the Prochlorococcus strains/ecotypes, whereas the other subclade mainly comprised various marine Synechococcus strains, except for the branching of Prochlorococcus marinus MIT9303 and MIT9313 within them. In clade B, a number of subclades including those corresponding to Nostoc/Anabaena (order Nostocales) and Synechocystis/Crocosphaera (order Chroococcales) were resolved with high statistical support (i.e. supported by >90% puzzled quartets or bootstrap samples). The branching position of Synechococcus elongatus was uncertain in these trees; in the ML tree (Supplementary Fig. S1), it branched in between clades B and C, whereas in the NJ tree (Fig. 1), it branched with clade B, although this association lacked statistical support.

An NJ tree for these cyanobacteria was also constructed based on 16S rRNA gene sequences (Fig. 2). This tree consisted of only two main clades, one consisting of the clade A and B cyanobacteria as well as Synechococcus elongatus, whilst the clade C species/strains/isolates comprised the other clade. In contrast to the tree based on the concatenated protein sequences (Fig. 1), the deep branches in the 16S rRNA gene-based tree, including that indicating the branching of Thermosynechococcus elongatus with clade A, were not supported statistically. These trees provide us with a framework for understanding the evolutionary significances of various conserved indels identified in this work.
Fig. 1. NJ distance tree for various cyanobacteria whose genomes are now available, based on concatenated sequences for 45 conserved proteins. The tree was rooted by using Bacillus subtilis and Staphylococcus aureus sequences. Numbers at nodes indicate bootstrap scores out of 1000. Bar, 0.1 substitutions per site.

Fig. 2. NJ distance tree for the same group of cyanobacteria as shown in Fig. 1, based on 16S rRNA gene sequences. The tree was bootstrapped 1000 times and numbers at nodes indicate percentage bootstrap scores. Bar, 0.1 substitutions per site.
Usefulness of signature indels (synapomorphies) for taxonomic/evolutionary studies

The shared derived characters that are unique to particular groups or clades of organisms (i.e., synapomorphies) provide an important means of identifying various monophyletic clades and also for understanding how these clades are related to each other. In the past, this approach has largely been employed by using morphology and other observable traits (Sneth & Sokal, 1973; Sneth, 2001). However, such traits are often either plesiomorphic (i.e. a particular character is not limited to a given group) or exhibit homoplasy (the derived character state has evolved independently in the given group of organisms), limiting their utility as phylogenetic or taxonomic markers. In recent years, the availability of genome sequences has led to the discovery of important molecular characteristics that are uniquely shared by different groups of organisms and provide important means for their grouping and for evolutionary studies. The conserved indels of defined lengths that are present in gene/protein sequences at specific positions, and which are uniquely shared by particular groups of organisms, form an important class of such characters (synapomorphies) and have proven very useful in clarifying many important evolutionary relationships (Rivera & Lake, 1992; Baldauf & Palmer, 1993; Gupta, 1998; Rokas & Holland, 2000). When a conserved indel of defined length and sequence is found specifically in a particular group of species, then the simplest and most parsimonious explanation for this observation is that these species shared a common ancestor, in which the rare genetic change responsible for this indel first occurred and was then passed on to various descendants (Rivera & Lake, 1992; Gupta, 1998; Rokas & Holland, 2000). However, such changes can also occur due to lateral gene transfers between or among species (Boucher et al., 2003; Zhaxybayeva et al., 2006); hence, it is important to exclude this possibility by comparing the inference obtained by the signature approach with that derived from other independent means, such as phylogenetic or phylogenomic methods. I describe below many conserved indels or molecular synapomorphies that are distinctive characteristics of various clades of cyanobacteria.

Conserved indels that are specific for cyanobacteria and their deepest-branching clades

Previously, we described 14 conserved indels in 10 proteins that are specific for all cyanobacterial homologues (Gupta et al., 2003). Since that time, sequence information for many cyanobacteria and a large number of other bacteria has become available. Hence, it was of importance to check the species specificities of these signatures. Of these 14 indels, 13 (including the 6, 7 and 28 aa inserts in the DNA helicase II protein, 14 aa insert in ADP-glucose pyrophosphorylase, 3 aa insert in FtsH protein, 11–13 aa insert in pythoene synthase, 5 aa insert in EF-Tu, 2 and 7 aa deletions in the ribosomal S1 protein, 2 aa insert in SecA protein, 6 aa insert and 1 aa deletion in inosine-5′-monophosphate dehydrogenase and 1 aa insert in the major sigma factor) were present in all sequenced (available) cyanobacteria, but not found in any other bacteria (results not shown). In this work, I have identified two additional conserved indels, viz. a 3 aa deletion in carbamoylphosphate synthase (Supplementary Fig. S2) and a 3 aa insert in prolyl-tRNA synthetase (Supplementary Fig. S3), that are also specific for cyanobacteria and are not found in any other bacteria. In prolyl-tRNA synthetase, a 2 aa insert is also found in various chlamydial species in the same position where the 3 aa insert is present in cyanobacteria. Earlier studies have indicated the presence of chlamydia-like genes in green plants and algae, but not in cyanobacteria (Horn et al., 2004; Huang & Gogarten, 2007; Moustafa et al., 2008). In a phylogenetic tree for PrRS sequences, the chlamydiae and cyanobacteria do not group together (results not shown), indicating that the inserts in these two groups, which are of different lengths, have originated independently. Because of their cyanobacterial specificities, the genetic changes responsible for these molecular signatures probably occurred in a common ancestor of cyanobacteria, as indicated in Fig. 3.

In contrast to the above indels, the 18 aa insert in DNA polymerase I (Pol I), identified previously (Gupta et al., 2003), was present in all other cyanobacteria except those corresponding to clade A (Fig. 4). This indel was also absent in various other bacteria, indicating that this synapomorphy is an insert in other cyanobacteria that was introduced after the divergence of clade A (see Fig. 3). In addition to the Pol I signature, we have identified several other conserved indels that support a similar relationship. These include a 4–5 aa insert in the tryptophan synthase beta chain (Supplementary Fig. S4), a 4 aa insert in tryptophanyl-tRNA synthetase (TrpRS) (Supplementary Fig. S5) and a 2 aa insert in the DNA polymerase III gamma/tau subunits (encoded by the dnaX gene) (Supplementary Fig. S6). The conserved inserts in these proteins are commonly shared by all other cyanobacteria, but they are absent in the clade A species/strains as well as other bacteria. These signatures/synapomorphies support the view that the clone A species/strains constitute the deepest-branching lineage within cyanobacteria (Fig. 3), which is in accordance with the phylogenetic tree based upon our protein dataset (Fig. 1; Supplementary Fig. S1) and also with earlier phylogenomic studies based on other datasets of proteins (Honda et al., 1999; Turner et al., 1999; Sánchez-Baracaldo et al., 2005; Shi & Falkowski, 2008; Swingley et al., 2008a). Interestingly, the insert in the tryptophan synthase beta chain (Supplementary Fig. S4), which is absent in the clade A cyanobacteria but found in all other cyanobacteria, is also found in various plant and algal homologues. The shared presence of this synapomorphy in the plant/algal homologues and various cyanobacteria, except those from clade A, provides evidence that the plant/algal homologues of this protein are derived from a cyanobacterium not belonging to clade A.
We have also identified a 15 aa insert in a highly conserved region of the protein synthesis elongation factor-G (EF-G) that is found only in the clade A cyanobacteria (Fig. 5). This synapomorphy provides a molecular marker for the clade A species/strains (subclass Gloeobacterophycidae) and, in conjunction with other signatures described above (Fig. 3), provides evidence that this group is distinct from all other cyanobacteria. A number of conserved inserts in other proteins are specifically present in the two Synechococcus strains JA-3-3Ab and JA2-3-B'a that are part of clade A. These include a 12 aa insert in the protein serine hydroxymethyltransferase (GlyA), an 8 aa insert in the RNA polymerase beta subunit (RpoB), a 6 aa insert in the ribosomal L2 protein and a 2 aa insert in the prolyl-tRNA synthetase (Supplementary Figs S7–S10).

**Signature indels for the clade C cyanobacteria**

As indicated earlier, clade C is composed entirely of different strains/isolates and ecotypes of Prochlorococcus and Synechococcus. Of these, Synechococcus spp. are ubiquitous in different aquatic environments, including estuarine, coastal and offshore waters (Palenik et al., 2006). In contrast, Prochlorococcus spp. are found mainly in warm, oligotrophic oceanic settings (Rocap et al., 2003). We have identified a number of conserved indels in widely distributed proteins that are specific for the clade C species/strains (subclass Synechococcophycidae) and also for one of its subgroups. The signatures that are specific for clade C include a 3 aa insert in the RNA polymerase beta subunit (RpoB; Fig. 6), a 2 aa insert in the protein KsgA (Supplementary Fig. S11) that carries out essential dimethylation of the 16S rRNA, a 6 aa insert in the tyrosyl-tRNA synthetase (Supplementary Fig. S12), a 2 aa insert in the tRNA (guanine-N1-)-methyltransferase (Supplementary Fig. S13) and a 1 aa insert in the RNA polymerase beta subunit (RpoC; Supplementary Fig. S14). The inserts in all of these proteins are present in all of the clade C species/strains (sequences for some closely related strains of Prochlorococcus and Synechococcus strains are not shown in these alignments), but they are not found in other cyanobacteria, including Acaryochloris marina, which has also been suggested to belong to the subclass Synechococcophycidae (Hoffmann et al., 2005). Another prominent signature for clade C (a 12 aa insert) is found in Pol I (Supplementary Fig. S15), which, as indicated...
previously, also contains a large insert that is specific for various cyanobacteria except those from clade A (Fig. 4). In contrast to the other clade C-specific signatures, the insert in Pol I is lacking in *Synechococcus* WH5701, which lies in the middle of the clade C species/strains. To account for the absence of the Pol I insert in *Synechococcus* WH5701, one must postulate that either the Pol I gene in this cyanobacterium was acquired laterally from some other bacterium lacking the insert, or that the insert in this case has somehow been deleted/lost. Based upon the available information, we are unable to distinguish between these possibilities. Of the inserts that are specific for the clade C cyanobacteria, the homologues of RpoB are also present in various plastids and green plants, and this insert is not found in any of them (Fig. 6). Likewise, the plant homologues of the protein tRNA (guanine-N1-)-methyltransferase also lack the insert in this protein (Supplementary Fig. S13). These observations suggest that the plastid/plant homologues of these proteins have not originated from clade C cyanobacteria.

Within clade C, among the *Prochlorococcus* isolates, two phylogenetically and physiologically different subgroups or ecotypes (high-B/A and low-B/A) have been identified (Moore et al., 1998; Rocap et al., 2002). The strains from these two ecotypes differ in terms of the relative ratios of chlorophyll b and a2 in their light-harvesting systems. Strains from these two subgroups also differ in their ability to grow at different light intensities and copper toxicities and to use nitrite or nitrate as nitrogen sources (Ferris & Palenik, 1998; Rocap et al., 2002). In the phylogenetic trees shown in Figs 1 and 2, the low-B/A *Prochlorococcus* isolates formed a distinct subclade within clade C that was separated from all other clade C species/strains by a long branch length and a 100 % bootstrap score. We have identified two conserved inserts that are specific for the plastid/plant homologues of these proteins have not originated from clade C cyanobacteria.

Within clade C, among the *Prochlorococcus* isolates, two phylogenetically and physiologically different subgroups or ecotypes (high-B/A and low-B/A) have been identified (Moore et al., 1998; Rocap et al., 2002). The strains from these two ecotypes differ in terms of the relative ratios of chlorophyll b and a2 in their light-harvesting systems. Strains from these two subgroups also differ in their ability to grow at different light intensities and copper toxicities and to use nitrite or nitrate as nitrogen sources (Ferris & Palenik, 1998; Rocap et al., 2002). In the phylogenetic trees shown in Figs 1 and 2, the low-B/A *Prochlorococcus* isolates formed a distinct subclade within clade C that was separated from all other clade C species/strains by a long branch length and a 100 % bootstrap score. We have identified two conserved inserts that are specific for the
low-B/A subgroup of *Prochlorococcus marinus* isolates. In the protein leucyl-tRNA synthetase, which is essential for protein synthesis, a 5 aa deletion is present that is specific for the various low-B/A subgroup *Prochlorococcus* strains (MIT9515, CCMP1986, MIT9312, MIT9215, MIT9301 and AS9601), but not found in any other cyanobacteria (Supplementary Fig. S16). Another 1 aa insert in a conserved region of the signal-recognition particle protein Ffh or SRP54 is also specific for the low-B/A subgroup of *Prochlorococcus* strains (Supplementary Fig. S17).

Additionally, in the TrpRS signature shown in Supplementary Fig. S5, all of the low-B/A subgroup *Prochlorococcus marinus* strains contained a 3 aa insert, whilst a 4 aa insert was found in other cyanobacteria. These signature indels provide evidence that the low-B/A subgroup of *Prochlorococcus marinus* strains is phylogenetically and molecularly distinct from all other clade C species/strains. The evolutionary stages where genetic changes responsible for these signature sequences probably occurred are indicated in Fig. 7.

**Conserved indels that are specific for the clade B cyanobacteria**

A number of conserved indels have been identified that are specific for the clade B cyanobacteria, which encompass the majority of known cyanobacteria (Honda *et al*., 1999; Turner *et al*., 1999; Castenholz, 2001; Wilmotte & Herdman, 2001; Sánchez-Baracaldo *et al*., 2005; Swingley *et al*., 2008a). Within clade B, heterocyst-forming cyanobacteria form a monophyletic group (subclass Nostocophycidae) (Turner *et al*., 1999; Adams, 2000; Wilmotte & Herdman, 2001; Hoffmann, 2005; Rajaniemi *et al*., 2005). These heterocystous cyanobacteria include both *Nostocales* (non-branching and dividing always in a plane at a right angle to the long axis of the trichome) and *Stigonematales* (showing true branching) (Rippka *et al*., 1979; Wilmotte & Herdman, 2001; Rajaniemi *et al*., 2005). We have identified two conserved indels that are specific for these cyanobacteria. In the PetA protein, which is a precursor of the apocytochrome f, a 4 aa insert in a conserved region is present in various *Nostocales* species/strains (*Nostoc, Anabaena* and *Nodularia*) as well as in *Mastigocladus laminosus*, which belongs to the order *Stigonematales* (Fig. 8). The unique shared presence of this insert in these cyanobacteria provides further evidence that they form a monophyletic group. Another conserved insert (5 aa) specific for the *Nostocales* is found in ribosomal protein S3 (Supplementary Fig. S18). The sequence information for this protein is not available for *Stigonematales*. In ribosomal protein S3, in the same position where the 5 aa insert is found in the *Nostocales*, an
8 aa insert is present in many green plants, but not in any of the plastid homologues (Supplementary Fig. S19). The different lengths and sequences of these inserts indicate that they are of independent origin.

Cyanobacteria such as *Synechocystis*, *Microcystis*, *Crocosphaera* and *Cyanothece*, etc., which belong to the order *Chroococcales*, form another well-defined clade in phylogenetic trees (Figs 1 and 2) (Honda et al., 1999; Turner et al., 1999; Sánchez-Baracaldo et al., 2005; Shi & Falkowski, 2008; Swingley et al., 2008a). We have identified a 1 aa insert in a highly conserved region of the RecA protein that is commonly shared by all of these cyanobacteria (Fig. 9). This insert is also present in *Synechococcus PCC7002*, which branches with this clade in the phylogenetic trees (Fig. 2) (Turner et al., 1999; Sánchez-Baracaldo et al., 2005).

In phylogenetic trees based on different datasets, *Trichodesmium erythraeum* generally branches in between the *Nostocales* and the *Chroococcales*, often as an outgroup of the *Nostocales* clade (Figs 1 and 2) (Honda et al., 1999; Turner et al., 1999; Sánchez-Baracaldo et al., 2005; Shi & Falkowski, 2008; Swingley et al., 2008a). We have identified a 4 aa deletion in translation-initiation factor IF-2 that, within cyanobacteria, is uniquely present in various *Nostocales*, as well as in *Trichodesmium erythraeum* and *Lyngbya PCC8106* (Supplementary Fig. S19), both of which belong to the order *Oscillatoriaceae*. The shared presence of this signature in the *Nostocales* and *Oscillatoriaceae* provides further evidence that these two orders of cyanobacteria are close relatives, as also seen in phylogenetic trees (Figs 1 and 2).

Three other large signature indels provide evidence that cyanobacteria from the orders *Nostocales*, *Oscillatoriaceae* and *Chroococcales* shared a common ancestor, exclusive of other cyanobacteria for which sequence information is presently available (Fig. 7). In the DnaE protein, a 19 aa insert is present specifically in all available sequences from *Synechococcus PCC7002*.
these orders, but not in any other cyanobacteria or bacteria (Fig. 10). Likewise, in the protein GDP-mannose pyrophosphorylase, a 13 aa deletion is specifically present in these cyanobacteria, to the exclusion of all others (Supplementary Fig. S21). Another prominent indel of about 22–27 aa in the protein NAD(P)H-quinone oxidoreductase subunit D (Supplementary Fig. S22) is also specific for this clade of cyanobacteria. Most cyanobacteria contain at least two homologues of this protein (Melo et al., 2004) and this insert is present in only one of the two homologues from these groups of cyanobacteria, but not in other cyanobacteria. The evolutionary stages where genetic changes responsible for these conserved indels have probably occurred are indicated in Fig. 7.

DISCUSSION

In this work, I have used a combination of phylogenomic and signature sequence-based (i.e. phenetic) approaches to understand the evolutionary relationships among cyanobacteria. Phylogenetic trees were constructed based on concatenated sequences for 45 widely distributed proteins from the genomes of 34 sequenced cyanobacteria, as well as 16S rRNA gene sequences. In parallel, I have also identified 28 new conserved indels that are distinctive characteristics of various cyanobacterial clades that can be resolved in these phylogenetic trees. The identified conserved indels are found uniquely in all species/strains from the indicated groups of cyanobacteria, but they are not found in other cyanobacteria/bacteria. These results suggest strongly that the rare genetic changes responsible for these conserved indels occurred only once in a common ancestor of these clades and that their species-distribution patterns have not been affected by non-specific mechanisms such as lateral gene transfers (Boucher et al., 2003; Zhaxybayeva et al., 2006). These conserved indels thus provide potentially useful molecular synapomorphies for the identification and circumscription of a number of monophyletic clades within the cyanobacteria. The cyanobacterial clades that

**Fig. 7.** Interpretive diagram showing various signatures that are specific for the clade B and C cyanobacteria and the evolutionary stages where they were probably introduced. The signatures for the nodes marked ● are described in Fig. 3.
Hoffmann et al. (2005) proposed a new classification scheme for cyanobacteria that divides them into four subclasses. Of the clades that we can identify based upon the identified signature sequences, clade A corresponds to the subclass Gloeobacterophycidae (Hoffmann et al., 2005). The two Synechococcus strains (JA-3-3A and JA-2-3B) that group reliably with this clade both in phylogenetic trees and based upon different signature sequences have presumably been misidentified as Synechococcus and they are clearly indicated to be part of this clade. The clade comprising the Nostocales (plus Stigonematales) corresponds to the subclass Nostocophycidae, whereas clade C corresponds to the subclass Synechococcophycidae, with the exception that Acaryochloris is indicated to not be a part of this clade. The fourth proposed subclass of cyanobacteria, Oscillatoriophycidae, comprises the orders Oscillatoriales and Chroococcales. We have identified a conserved insert in the protein riboflavin synthase that is uniquely shared by the available sequences from these orders. Although this signature supports a grouping of these two orders in the subclass Oscillatoriophycidae (Hoffmann et al., 2005), the results from various phylogenetic studies (Figs 1 and 2) (Sánchez-Baracaldo et al., 2005; Swingley et al., 2008a), as well as a signature in the IF-2 protein (Supplementary Fig. S19), indicate that the order Oscillatoriales, represented in this work by Trichodesmium erythraeum and Lyngbya PCC8106, is related more closely to the Nostocales than to the Chroococcales. However, it should be noted that the filamentous non-heterocystous cyanobacteria such as those belonging to the order Oscillatoriales are a heterogeneous and possibly polyphyletic assemblage (Hoffmann et al., 2005). Although our analysis based on the genomes for Trichodesmium erythraeum and Lyngbya PCC8106 suggests an intermediate placement of the order Oscillatoriales between the orders Nostocales and Chroococcales (Fig. 7), it is possible that this placement could be a consequence of limited representation of taxa (only two genomes) from this order in our work. Hence, it is important to confirm this inference by using sequence data for various other members of the orders Oscillatoriales, Nostocales and Chroococcales.

In addition to these signatures, a number of other identified synapomorphies clarify certain evolutionary
relationships that are not resolved in phylogenetic trees. Species belonging to the orders Nostocales, Oscillatoriales and Chroococcales are indicated as late-diverging lineages within clade B of cyanobacteria (Honda et al., 1999; Turner et al., 1999; Sánchez-Baracaldo et al., 2005; Swingley et al., 2008a). However, a clade comprising these cyanobacteria is only weakly supported in some phylogenetic trees (Sánchez-Baracaldo et al., 2005; Swingley et al., 2008a). In the present work, we have identified three prominent signatures that are uniquely shared by various sequenced taxa from the orders Nostocales, Oscillatoriales and Chroococcales. These results provide strong suggestive evidence that species from these orders shared a common ancestor exclusive of all other cyanobacteria and that these signatures are reflecting a higher taxonomic clade within the cyanobacteria.

The evolutionary relationship among cyanobacteria was studied in this work based upon species/strains whose genomes have been sequenced (Table 1). About 60% of these genomes are from marine Prochlorococcus and Synechococcus strains/isolates. Thus, the present dataset does not truly reflect the relative abundance or genetic diversity of this phylum. Nonetheless, the available genomes cover many important groups/orders within cyanobacteria and they permit comparison of the results obtained by using the 16S rRNA gene with those obtained by other approaches. Some of the clades, such as those corresponding to the orders Nostocales, Chroococcales or clade C, are clearly distinguished in all of the trees (Figs 1 and 2) as well as by means of conserved indels in protein sequences. However, the deeper branching of the clade A cyanobacteria, whilst resolved with strong statistical support in the phylogenomic tree based on protein sequences (Fig. 1) and also based on conserved indels (Fig. 3), was not resolved in the 16S rRNA gene-based tree (Fig. 2). Importantly, on the basis of conserved indels, we were able to clearly distinguish not only all of the clades that can be resolved in both the 16S rRNA gene and the protein trees, but also a deeper-branching clade comprising the orders Nostocales, Chroococcales and Oscillatoriales (Fig. 7) that was not resolved by phylogenetic methods. These observations indicate that the results obtained by using the phylogenomic and signature-sequence approaches are, in general, in very good agreement with those suggested by 16S rRNA gene analysis, but these other methods are better suited for resolving deeper-branching relationships that are not resolved in 16S rRNA gene-based trees (Wilmotte & Herdman, 2001; Hoffmann et al., 2005).
16S rRNA gene-based trees currently provide the primary means for understanding microbial taxonomy and phylogeny. Currently, 16S rRNA gene sequences from 4500 cyanobacteria are available in the Ribosomal Database Project (RDP release 10; Maidak et al., 2001). However, with the large increase in the numbers of these sequences, the ability of 16S rRNA gene-based trees to resolve the relative branching positions of different taxa (particularly the deeper branches) has diminished greatly and, in some cases, been almost entirely lost (Ludwig & Schleifer, 1999; Wilmotte & Herdman, 2001; Hoffmann et al., 2005). This loss of phylogenetic clarity is in large part a consequence of many variables, such as differences in base compositions or evolutionary rates among various lineages, long branch-length effect, etc., that affect the branching of species in phylogenetic trees markedly, but are very difficult to control or correct for (Felsenstein, 2004; Delsuc et al., 2005). Hence, other stable characteristics that are consistent with the 16S rRNA gene-based as well as other phylogenetic approaches, but are affected minimally by these variables, are of particular importance for taxonomic and evolutionary studies (Rivera & Lake, 1992; Delwiche et al., 1995; Gupta, 1998, 2000; Rokas & Holland, 2000; Gupta et al., 2003; Skophammer et al., 2007; Griffiths & Gupta, 2007).

Conserved indels, because the genetic changes that give rise to them are of a highly specific nature, are less likely to occur independently in different taxa. Further, because these indels are present in highly conserved regions of the proteins, their presence or absence is not affected by differences in base composition or evolutionary rates among various lineages. Hence, such indels have provided generally reliable synapomorphic characteristics for identifying different monophyletic clades that are due to commonly shared ancestry (Rivera & Lake, 1992; Delwiche et al., 1995; Gupta, 1998, 2000; Gupta et al., 2003; Skophammer et al., 2007; Griffiths & Gupta, 2007).

In this work, we have also evaluated the reliability of many cyanobacterial signatures that were described in earlier work (Gupta et al., 2003). Of the 14 cyanobacteria-specific indels described previously when sequence data were available for only eight cyanobacteria, 13 are present in all sequenced cyanobacteria. One of these signatures (the insert in Pol I; Fig. 4) is absent only in clade A cyanobacteria for which no sequence information was available at that time. Similar results have been obtained for numerous other indels that are distinctive characteristics of other groups/phyla of bacteria (Gao & Gupta, 2003).

**Fig. 10.** Excerpt from a sequence alignment for the DnaE protein showing a 19 aa insert (boxed) that is specific for the sequenced species/stains from the orders *Nostocales*, *Oscillatoriales* and *Chroococcales*, but not found in any other cyanobacteria. This insert was probably introduced into a common ancestor of these orders, as indicated in Fig. 7.
These observations indicate strongly that the vast majority of these conserved indels retain their specificity as sequence information from new genomes or other sources becomes available and that many of these signatures will prove useful as reliable molecular markers for taxonomic and evolutionary studies. It should be emphasized that the placement of any species/strain into a given clade by using this approach is based upon several independent signatures that provide complementary information. Some of these signatures serve to exclude a given species/strain from particular groups or clades, whereas others point to its inclusion in more and more specific clades. The information provided by all of the signatures is generally internally highly consistent and only in exceptional cases is contradictory placement of a species in alternative position(s) indicated. For placement of any species or strain into a particular clade, in most cases, partial sequence information for only a few genes containing the diagnostic signature indels is sufficient. Because all of these signatures are flanked on both sides by conserved sequence regions, degenerate PCR primers for obtaining the requisite sequence information from other cyanobacteria could be designed readily. Hence, it should be of much interest to obtain sequence information for many of these signatures from a broad range of cyanobacteria, to confirm the usefulness and reliability of these molecular markers for taxonomic studies.

The present work also provides useful insights into the evolutionary relationship between cyanobacteria and plastids. The shared presence of several cyanobacteria-specific indels in the plastid/plant homologues (Gupta et al., 2003) has previously supported the view that plastids have originated from a cyanobacterial ancestor (Gray & Doolittle, 1982; Morden et al., 1999; Delwiche et al., 1995; Palmer & Delwiche, 1998; Gupta et al., 2003; Archibald, 2006; Stiller, 2007). In this work, the shared presence of the 5 aa insert in the tryptophan synthase beta subunit (Supplementary Fig. S4) in different green plants and algae and all cyanobacteria except those from clade A provides evidence that the plant/plastid homologues have not originated from clade A. The signature indels in several other proteins (RpoB, Fig. 6; RpoC, Supplementary Fig. S14) that are uniquely found in the Clade C homologues, but are absent in other cyanobacteria as well as plant/plastid homologues, provide evidence against the origin of plastids from clade C cyanobacteria. By exclusion, these results suggest that the plant/plastid homologues have probably originated from clade B cyanobacteria. A recent comparative genomic study has suggested that the plastid homologues are most similar to the order Nostocales (part of clade B) (Deusch et al., 2008). However, in our work, conserved inserts in the PetA (Fig. 8) and ribosomal S3 (Supplementary Fig. S18) proteins, which are specific for the order Nostocales, are not found in the corresponding plastid homologues, thereby questioning the validity of this inference. Hopefully, in future studies, additional signature sequences will be identified that will allow us to further pinpoint the cyanobacterial ancestor of the plastids.

**ACKNOWLEDGEMENTS**

This work was supported by a research grant from the Natural Science and Engineering Research Council of Canada. I thankfully acknowledge the technical assistance provided by Larissa Shamseer, Adeel Mahmood and Divya Wilson in the sequence-alignment studies and in the creation of some of the signature files.

**REFERENCES**


indels found in various proteins.

32

Mol Microbiol 55, 2401–2412.


Griffiths, E. & Gupta, R. S. (2001). The use of signature sequences in different proteins to determine the relative branching order of bacterial divisions: evidence that Fibrobacter diverged at a similar time to Chlamydia and the Cytophaga–Flavobacterium–Bacteroides division. Microbiology 147, 2611–2622.


