An overview of the genome of *Nostoc punctiforme*, a multicellular, symbiotic cyanobacterium.

John C. Meeks[1*], Jeff Elhai[2], Teresa Thiel[3], Malcolm Potts[4], Frank Larimer[5], Jane Lamerdin[6] Paul Predki[7] and Ronald Atlas[8]

Abstract

Nostoc punctiforme is a filamentous cyanobacterium with extensive phenotypic characteristics and a relatively large genome, approaching 10 Mb. The phenotypic characteristics include a photoautotrophic, diazotrophic mode of growth, but *N. punctiforme* is also facultatively heterotrophic; its vegetative cells have multiple developmental alternatives, including terminal differentiation into nitrogen-fixing heterocysts and transient differentiation into spore-like akinetes or motile filaments called hormogonia; and *N. punctiforme* has broad symbiotic competence with fungi and terrestrial plants, including bryophytes, gymnosperms and an angiosperm. The shotgun-sequencing phase of the *N. punctiforme* strain ATCC 29133 genome has been completed by the Joint Genome Institute. Annotation of an 8.9 Mb database yielded 7,432 open reading frames, 45% of which encode proteins with known or probable known function and 29% of which are unique to *N. punctiforme*. Comparative analysis of the sequence indicates a genome that is highly plastic and in a state of flux, with numerous insertion sequences and multilocus repeats, as well as genes encoding transposases and DNA modification enzymes. The sequence also reveals the presence of genes encoding putative proteins that collectively define almost all characteristics of cyanobacteria as a group. *N. punctiforme* has an extensive potential to sense and respond to environmental signals as reflected by the presence of more than 400 genes encoding sensor protein kinases, response regulators and other transcriptional factors. The signal transduction systems and any of the large number of unique genes may play essential roles in the cell differentiation and symbiotic interaction properties of *N. punctiforme*

Introduction: Physiological capabilities of *Nostoc punctiforme*

Nostoc punctiforme is a nitrogen fixing cyanobacterium, found predominantly in terrestrial habitats. *N. punctiforme* displays an extraordinarily wide range of vegetative cell developmental alternatives, physiological properties and ecological niches, including symbiotic associations with plants and fungi. The multiple phenotypic characteristics and growth habitats of *N. punctiforme* are indicative of the breadth of genetic information that is likely to be present in its genome. Completion of the shotgun sequencing phase of the *N. punctiforme* genome also indicates an exceptionally large microbial genome, approaching 10 Mb. Herein, we will summarize the genetic potential detectable in the currently unfinished genome of *N. punctiforme* that encompasses nearly all characteristics that define cyanobacteria as a group.

The cyanobacteria, or oxyphotobacteria (formerly called blue-green algae), are a very ancient group of microorganisms, with a morphological fossil record extending to approximately 3.0 to 3.5 billion years ago (Ga) and a chemical record to approximately 2.8 Ga (Des Marais 2000). All cyanobacteria are uniquely characterized as oxygen evolving photoautotrophic prokaryotes that employ chlorophyll *a* as the photochemically active pigment. Ancestors to extant cyanobacteria profoundly changed the biosphere in two ways: First, the photosynthetic production of oxygen slowly saturated the reactive chemicals in the immediate aquatic habitats (e.g. oxidation of iron to establish geological banded iron formations, between 3.0 and 2.0 Ga), before bringing the atmosphere to near its current oxygen content (Schopf 2000). Second, the formation of endosymbiotic associations with eukaryotic cells led to the evolution of chloroplast-containing algae and terrestrial plants (Douglas 1994), an event that occurred between 1.5 and 0.6 Ga, after free oxygen had become a significant selective force in the environment.

N. punctiforme is classified in the cyanobacterial order Nostocales, defined by an unbranched filamentous morphology and the ability to differentiate heterocysts (Fig. 1) (Castenholz and Waterbury 1989). Heterocysts are microoxic cells specialized for nitrogen fixation in an oxic environment (Wolk et al. 1994). The multiple phenotypic characteristics of *N. punctiforme* are listed in e 1. Many of the phenotypic characteristics of *N. punctiforme* are directly relevant to increasing its fitness for

[1]Section of Microbiology, University of California, Davis, CA 95616. [2]Department of Biology, University of Richmond, Richmond, VA 23173. [3]Department of Biology, University of Missouri, St. Louis, MO 63121. [4]Department of Biochemistry, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. [5]Computational Biology, Oak Ridge National Laboratory, Oak Ridge, TN 37831. [6]Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94551. [7]DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598. [8]Department of Biology, University of Louisville, Louisville, KY 40292. *Author for correspondence: telephone 530-752-3346, Fax 530-752-9014, email jcmeeks@ucdavis.edu

Table 1. Phenotypic characteristics of *Nostoc punctiforme*

1. Primarily an oxygenic photoautotrophic, diazotrophic mode of growth
2. Type II complementary chromatic adaptation by varying phycoerythrin content
3. Produces UV-absorbing compounds in response to UV light
4. Regulates acquisition of inorganic nutrients, especially nitrogen ($NH_4^+ > NO_3^- > N_2$)
5. Dark heterotrophic growth on sucrose, glucose or fructose
6. Multiple developmental alternatives exemplifying a complex life cycle (Fig. 1)
7. Buoyant and gliding motility with photo- and chemotactic responses
8. Broad symbiotic competence with fungi and terrestrial plants

photosynthesis. *N. punctiforme* synthesizes the phycobiliproteins phycoerythrin (PE), phycocyanin (PC) and allophycocyanin (APC), which are assembled into a multiprotein complex, the phycobilisome. Except for the chlorophyll *a* and *b* containing oxychlorobacteria (prochlorophytes that phylogenetically cluster within the division cyanobacteria; Hess et al., this volume), APC and PC are found in all cyanobacteria, while the presence of PE is variable with no obvious taxonomic correlation. Cyanobacterial strains containing PE may undergo a process of complementary chromatic adaptation (CCA) in response to light quality. *N. punctiforme* is categorized as having Type II CCA in that the synthesis of PE, but not PC, is controlled by the presence or absence of green light; this is in contrast to Type III CCA where synthesis of PE and PC are controlled by the reciprocal presence or absence of green and red light (Tandeau de Marsac 1977). CCA is of a clear advantage to light harvesting by cyanobacteria in competitive habitats, such as microbial mats and soils beneath forest canopies, where the spectral quality of the light may frequently change.

Cyanobacteria that grow in full light as unshaded mats and on soils are subjected to high incidences of UV light and many respond by the synthesis of UV light absorbing compounds. *N. punctiforme* synthesizes two classes of UV light absorbing compounds; scytonemin (Hunsucker et al. 2001) and microsporin-like amino acids (F. Garcia-Pichel, personal communication). Although cyanobacteria are also well known for repair of UV-induced DNA damage (Levine and Thiel 1987), the presence of UV light absorbing compounds increases the chances of survival and continued photosynthesis in high light exposed habitats.

N. punctiforme can utilize various inorganic and organic nitrogen sources for growth. The inorganic nitrogen sources are assimilated in the hierarchical order of $NH_4^+ > NO_3^- > N_2$. Growth on N_2 requires the differentiation of heterocysts (see below) and heterocyst differentiation is repressed by the presence of NO_3^- or NH_4^+ in *N. punctiforme* (Campbell and Meeks 1992) and essentially all heterocyst forming cyanobacteria (Wolk et al. 1994). Nitrate assimilation is repressed by the presence of NH_4^+ in all cyanobacteria examined (Flores and Herrero 1994). Flexibility in utilization of various nitrogen sources for growth allows *N. punctiforme* to colonize and compete as a phototroph in illuminated habitats, irrespective of the specific nitrogen source.

N. punctiforme vegetative cells have three developmental fates that are dependent on the environmental growth conditions. Heterocyst formation is one of the developmental alternatives. These nitrogen-fixing cells terminally differentiate in response to a limitation in combined nitrogen and appear at a frequency of 8-9% of the total cells in a well spaced pattern within the filament (Fig. 1). Heterocysts are highly modified in their metabolism, maintaining the microoxic cellular environment essential for nitrogen fixation by, in part, elimination of the oxygenic photosynthetic reactions and conversion to a heterotrophic metabolic mode with a high respiration rate (Wolk et al. 1994). Heterocyst differentiation and maintenance is variously estimated to involve no more than 140 (Wolk 2000) to about 1,000 (Lynn et al.1986) genes.

Under conditions of cellular energy limitation imposed by, for example, phosphate limitation, all cells transiently differentiate into spores (Fig. 1), called akinetes in cyanobacteria. These structures remain viable for hundreds of years under desiccated conditions prior to germination into vegetative filaments (Adams and Duggan 1999). Akinetes are speculated to be progenitors of heterocysts (Wolk et al. 1994); there is little genetic information

(Picture not available)

Figure 1. Photomicrograph of *Nostoc punctiforme* filaments showing the various developmental states. Heterocysts, akinetes and hormogonium filaments are indicated. The image is phase contrast; vegetative cells are typically 5-6 μm in diameter, heterocysts 6-10 μm and akinetes 10-20 μm.

available on their differentiation or germination.

In response to stress signals, all cells within a filament may divide in the absence of biomass increase or DNA replication to transiently form motile filaments called hormogonia (Fig. 1). These gliding filaments serve as propagules in the colonization of new portions of a habitat. Motile hormogonia express photo- and chemotactic behavior. Many hormogonia, including those of *N. punctiforme* (Rippka and Herdman 1992), possess an additional motility system: gas vesicles provide buoyancy in soil or aquatic water columns. Hormogonia return to vegetative filaments by differentiation of heterocysts, resumption of biomass increase and initiation of DNA replication (Meeks 1998).

The transition of vegetative filaments through akinete and hormogonium states is reflective of a complex life cycle in *N. punctiforme*. The patterned spacing of heterocysts in filaments and the source-sink relationship between vegetative cells (supplier of reduced carbon and sinks for reduced nitrogen) and heterocysts (supplier of reduced nitrogen and sinks of reduced carbon) reflect extensive cell-cell communication and suggests that *N. punctiforme*, and related heterocyst-forming cyanobacteria, are also by definition multicellular organisms.

N. punctiforme is amongst a limited number of cyanobacteria that can grow in continual darkness as a respiratory heterotroph when supplied with sucrose, glucose or fructose, although the rate is less than half of the photoautotrophic rate (Summers et al. 1995). The capacity for prolonged heterotrophic growth by *N. punctiforme* correlates with induced synthesis of glucose-6-phosphate dehydrogenase, the initial enzyme of the oxidative pentose phosphate pathway (OPP). The OPP, rather than glycolysis, is the primary route of carbon catabolism in cyanobacteria (Summers et al. 1995). The heterotrophic capacity of *N. punctiforme* also is consistent with its ability to grow in symbiotic association with plants.

N. punctiforme has broad symbiotic competence with fungi and plants. Strain ATCC 29133 (PCC 73102) was isolated as a symbiont from a coralloid root of the gymnosperm cycad *Macrozamia* sp. (Rippka and Herdman 1992). Cultured *N. punctiforme* ATCC 29133 has been experimentally documented to establish a nitrogen-fixing symbiosis with the bryophyte hornwort *Anthoceros punctatus* (Enderlin and Meeks,1983) and the angiosperm *Gunnera* spp. (Johansson and Bergman 1994). *N. punctiforme* is also identified as the intracellular symbiont of the unique mycorrhizal fungus *Geosiphon pyriforme* (Mollenhauer et al. 1996). The physiology of symbiotic *Nostoc* species is profoundly changed by their interaction with plants: growth and photosynthetic capabilities are diminished, the rate of nitrogen fixation is increased and heterotrophic metabolism in support of nitrogen fixation is enhanced. The plant partners have been shown to influence the differentiation and behavior of *N. punctiforme* hormogonia through molecular signals (Campbell and Meeks 1989, Cohen and Meeks 1997) and are presumed to similarly affect heterocyst differentiation and metabolism within the symbiosis (Campbell and Meeks 1992; Meeks 1998). The broad symbiotic competence of *N. punctiforme* and a corresponding lack of plant specificity for a single *Nostoc* strain imply that very diverse plants have adapted mechanisms to manipulate different heterocyst-forming cyanobacteria in the establishment of NH_4^+ producers for their growth. A planned functional genomic analysis of *N. punctiforme* will undoubtedly reveal genes uniquely involved in microbe-plant interactions leading to a stable nitrogen-fixing association.

We will present the overall characteristics and notable properties of the *N. punctiforme* genome in this review, with an emphasis on its photoautotrophic and diazotrophic phenotype. The *N. punctiforme* genome characteristics will broadly be compared to those of some published bacterial genomes, to the finished genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC 6803 (http://www.kazusa.or.jp/cyano) and the unfinished genomes of the closely related heterocyst-forming *Anabaena* sp. strain PCC 7120 (Kazusa) and the unicellular marine cyanobacteria *Synechococcus* sp. strain WH 8102 and *Prochlorococcus marinus* MED4 (both at http://www.jgi.doe.gov).

Methods

A whole genome shotgun sequencing strategy, pioneered by Fleischmann et al. (1995), was used to complete the sequence of the approximate 9.5 Mb genome of *N. punctiforme* strain ATCC 29133. DNA preparation protocols are posted at http://www.jgi.doe.gov/ under "Production protocols". Sequencing was done using Molecular Dynamics MegaBACE and Applied Biosystems ABI Prism 3700 instruments. Shotgun sequence data were assembled with PHRAP and through "auto-finishing" using software written by Matt Nolan (JGI/Lawrence Livermore National Laboratory) and David Gordon (University of Washington).

Assembly of a 9,125,999 bp database yielded 662 contigs; 320 contigs reflecting greater than 8 reads and 8,941,326 bp were annotated. The three gene modeling programs utilized to define open reading frames (ORF) were Critica, Glimmer and Generation. The results of the three gene-callers were combined and a BLASTP search of the translations versus the total gene sequence database (NR) was conducted. The alignment of the N-terminus of each gene model versus the best NR match was used to pick a preferred gene model and translation start point. If no BLAST match was returned, the longest model was retained. Gene models that overlapped by greater than 10% of their length were flagged, giving preference to genes with a BLAST match. The revised gene/protein set was searched against the KEGG GENES, Pfam, PROSITE, PRINTS, ProDom and COGS databases, in addition to BLASTP versus NR. In pair-wise similarity searches of gene-model translations, the BLAST E-value

Table 2. Numerical properties of the genome of *Nostoc punctiforme* strain ATCC 29133

Current genome size	~9,500,000 bases (11.4 x sequencing coverage; February, 2001)
Size of annotated sequence	8,941,326 bases (8 x sequencing coverage; 94% of genome; June 6, 2000)
Preliminary analyses	7,432 protein encoding ORFs identified
Relative to total database	5,314 of the ORFs (71% of the total) can be associated with a previously recognized ORF
	3,328 of the recognized ORFs (45% of the total) encode proteins with known or probable known function
	1,986 of the recognized ORFs (27% of the total) encode conserved hypothetical and hypothetical proteins with no known function
	2,164 of the ORFs (29% of the total) can NOT be associated with a previously recognized ORF

threshold was set to 1e-05. The initial Pfam threshold was set at 1e-05. Enzyme catalog references were obtained by parsing the BLAST versus total database. Putative genes were organized into functional categories based on KEGG categories and COGs hierarchies. Additional analyses are based on sequence comparisons to Cyanobase, Genbank and Swissprot. Analysis of oligonucleotide frequencies was performed using locally written software.

Broad overview of the genome

The overall numerical analysis of the *N. punctiforme* ATCC 29133 genome is given in Table 2. An 11.4 x sequence coverage database of between 9.25 and 9.5 Mb that is currently being analyzed implies that the annotated database in Table 2 reflects at least 94% of the actual genome. Thus, there are between 7,432 and less than 8,000 protein encoding ORFs in the complete genome. This is amongst the largest of the microbial genomes being sequenced, including the developmentally complex heterotrophic bacteria *Streptomyces coelicolor* and *Myxococcus xanthus* and the budding yeast *Saccharomyces cerevisiae*. A plot of the frequency of *N. punctiforme* ORFs as a function of predicted molecular mass yielded an average protein size of 35.9 kDa. Because the genome is currently unfinished, we have not attempted detailed analysis of gene location, orientation or operon structure. The annotated sequence is available at http://www.jgi.doe.gov/tempweb/JGI_microbial/html/nostoc/nostoc_homepage.html.

Seventy-one percent of the *N. punctiforme* ORFs can be associated with a previously recognized ORF. The presence of 3,328 ORFs (45%) that encode proteins with known or probable known function implies the likely occurrence of multiple gene families with core metabolic function. The observation that about 29% of the putative ORFs in the *N. punctiforme* genome have no significant similarity to sequences in the current database is similar to results in other microbial genomes, such as *Pseudomonas aeruginosa* (Stover et al. 2000).

Broad comparisons to Synechocystis *PCC 6803 and* Anabaena *PCC 7120*

Synechocystis PCC 6803 has a 3.57 Mb genome and contains 3,215 protein-encoding ORFs; 1,521 of those ORFs encode proteins with known or probable known function, while 1,694 ORFs encode conserved hypothetical or hypothetical proteins. BLAST analysis showed that 3,965 *N. punctiforme* ORFs have significant similarity in the *Synechocystis* PCC 6803 genome, while 2,547 *Synechocystis* PCC 6803 ORFs are similar to *N. punctiforme* ORFs. These results imply that the large *N. punctiforme* genome is not an exact multiple of the 2.7-fold smaller *Synechocystis* PCC 6803 genome; however, the 1,418 ORFs in excess of those of *Synechocystis* PCC 6803 that find similarity in *N. punctiforme* must represent multiple members of gene families that could have evolved from simple gene families in a common ancestor to both extant organisms. This analysis also showed that 668 of the *Synechocystis* PCC 6803 ORFs are not present in *N. punctiforme*, which indicates that, like other microorganisms, *Synechocystis* PCC 6803 has its own unique genes.

The *Anabaena* PCC 7120 genome, at 7.13 Mb with 5,610 ORFs, is less than 75% of the size of *N. punctiforme*. The reciprocal BLAST results indicate that 5,431 of the *N. punctiforme* ORFs are present in *Anabaena* PCC 7120 and 4,814 (86%) of the *Anabaena* PCC 7120 ORFs are present in *N. punctiforme*. Since 617 more *N. punctiforme* ORFs find similarity in *Anabaena* PCC 7120 than the reciprocal, it appears that *N. punctiforme* contains multiple copies of some *Anabaena* PCC 7120 ORFs. *Anabaena* PCC 7120, like *Synechocystis* PCC 6803, contains a similar number of ORFs (797) that appear unique relative to *N. punctiforme*. Of the 1,283 ORFs of *Anabaena* PCC 7120 that are unique in the database, 486 find similarity in the unique ORFs in the *N. punctiforme* genome. This observation has two implications: First, these shared ORFs may encode proteins involved in common phenotypic characteristics, such as heterocyst differentiation. Second, the actual number of unique ORFs in *N. punctiforme* relative to the

Table 3. Functional categories of predicted known or probable known genes

Functional category	Number of putative genes[a]	Percent of total genes[b]
1. Energy metabolism [photosynthetic & respiratory electron transport, ATP synthesis]	98	1.32
2. Inorganic nutrients [transport & metabolism of C, N, S, P, Fe, H & other ions]	204	2.74
3. Organic carbon [transport & metabolism of glycogen, sucrose, hexoses, pentoses & carboxylates]	200	2.69
4. Metabolic dehydrogenases, oxidoreductases, monooxygenases	134	1.80
5. Amino acids [transport & metabolism, including cyanophycin]	272	3.66
6. Nucleotides [transport & metabolism]	59	0.79
7. Lipids, fatty acids, polyketides & cyclic peptides	137	1.84
8. Coenzymes, vitamins, porphyrins & bilins	113	1.52
9. Cell secretion & chemotaxis	61	0.82
10. Cell envelope synthesis, cell division & chromosome segregation	278	3.74
11. DNA repair, replication & recombination, including transposases	271	3.65
12. Transcription [RNA polymerase & transcripitonal regulators]	82	1.10
13. Translation [tRNA synthetases, ribosomes & initiation]	144	1.93
14. Signal transduction mechanisms [protein kinases, response regulators & cAMP]	373	5.02
15. Protein modification [including molecular chaperones, glutathione, thioredoxin & proteases]	152	2.05
16. Unassigned probable enzymes & structural proteins	750	10.09

[a]The numbers of genes in each category are provisional; many have been analyzed for functional sites, completeness of sequence and accuracy in the gene identity, but the numbers will reflect the mistakes and omissions inevitable from an automated annotation.

[b]The percentage is based on the current total of 7,432 ORFs.

total database, now including *Anabaena* PCC 7120, reduces to 1,578 or 23% of the genome.

Functional categories of N. punctiforme *ORFs*

Based on automated annotation results, the 3,328 ORFs that encode known or probable known functional proteins were organized into categories that are most relevant to the *N. punctiforme* life style (Table 3). The numbers of genes involved in energy metabolism, nucleotide metabolism, coenzymes and pyrroles, and protein synthesis are similar to those of *Synechocystis* PCC 6803 and the heterotrophic bacteria *Bacillus subtilis* (Kunst et al. 1997), *Escherichia coli* (Blattner et al. 1997) and *P. aeruginosa* (Stover et al. 2000). *N. punctiforme* has an unusually large number of genes encoding alcohol and aldehyde dehydrogenases, oxidoreductases and putative F420 monooxygenases. These genes would be categorized within energy metabolism in heterotrophic organisms, but it is unlikely that their role is to provide reductant for respiratory electron transport in *N. punctiforme*; thus, they were placed in a separate category. Their metabolic roles are unknown.

Consistent with its ability to grow in the light on CO_2 and simple salts, *N. punctiforme* has a large number of genes for the acquisition and metabolism of inorganic nutrients. Included in this category are 91 genes encoding ion uptake and efflux transport components. *N. punctiforme* also has a comparatively large number of genes involved in lipid, fatty acid and complex polyketide synthesis, cell envelope synthesis, DNA metabolism, signal transduction and protein modification. The lipid, fatty acid and complex polyketide category contains two groups of genes encoding multidomain proteins. One group of polyketide synthases is involved in the synthesis of the unique heterocyst glycolipids. The other group is involved in the synthesis of cyclic peptide secondary metabolites, which will be discussed later. The cell envelope/division category includes 74 glycosyl transferases and 23 ATPases involved in chromosome partitioning that are infrequently detected in other bacteria. The DNA metabolism category includes approximately 150 putative transposases (see below); only *Synechocystis* PCC 6803 has a comparable number of putative transposase genes. The signal transduction category includes 255 genes encoding sensory histidine kinase and response regulator proteins and 55 serine/threonine protein kinases. These genes are present in high numbers relative to other bacteria and will be discussed later. The protein modification category includes, in addition to various chaperonins, multiple genes encoding various proteases, ATPases, and proteins for the synthesis and metabolism of thioredoxin and glutathione.

N. punctiforme has many genes involved in amino acid transport (56) and metabolism (216), whose total is

Table 4: Base frequencies in cyanobacterial genomes[a]

	N. punctiforme	*Anabaena* PCC 7120	*Synechocystis* PCC 6803	*P. marinus* MED4	*Synechococcus* WH8102
Estimated size	9.5 Mb	7.1 Mb	3.6 Mb	1.7 Mb	2.4 Mb
Mol %GC	41.5	41.2	47.7	30.9	58.5
Underrepresented dinucleotides[b]	CG 0.81 TA 0.81	CG 0.78 TA 0.84	CG 0.75 TA 0.75	CG 0.51 TA 0.79 GT/AC 0.72	CG 0.87 TA 0.43
Overrepresented dinucleotides[b]	GC 1.24		GG/CC 1.36 AA/TT 1.32	GG/CC 1.28	TG/CA 1.26

[a]Analysis uses complete sequence of *Synechocystis* PCC 6803 and partial sequences available for *N. punctiforme* (June 2000), *Anabaena* PCC 7120, *P. marinus* MED4, and *Synechococcus* WH8102.

[b]Under- and overrepresentation is given as the ratio (termed ρ) of the dinucleotide frequency expected from the %GC to the actual dinucleotide frequency (Karlin et al. 1997). All dinucleotides with ρ values less than 0.78 or greater than 1.22 are given, as well as values for CG and TA dinucleotides.

equivalent to those in *B. subtilis* and nearly twice, or more, those in *P. aeruginosa* and *Synechocystis* PCC 6803, respectively. Based on the *N. punctiforme* nutritionally independent life style, such a robust amino acid transport capacity was unanticipated. The number of *N. punctiforme* genes involved with organic carbon metabolism is less than half those of *B. subtilis*, but 1.5 times higher than those identified in *Synehcocystis* PCC 6803. Only sucrose, glucose and fructose, which can be catabolized through the oxidative pentose phosphate pathway, support heterotrophic growth of *N. punctiforme*; there is no evidence for diverse pathways of carbon metabolism to generate hexoses from other compounds or polymers. Transport of the hexoses to support growth appears more than adequate since *N. punctiforme* has 52 genes encoding organic carbon transport proteins, 75% of which appear to be for simple sugars. The role of the additional hexose transport genes may be for other than catabolic substrate supply.

N. punctiforme has distinctly fewer genes encoding transcriptional regulatory factors and cell secretion and chemotaxis proteins than does *P. aeuroginosa* which contains 448 and 191 ORFs in these two respective categories. Nevertheless, the presence of 13 alternative group 2 sigma subunits of RNA polymerase and 46 transcriptional regulators, in addition to 61 response regulators with DNA binding motifs, implies substantial capacity for differential gene expression in *N. punctiforme*, which is consistent with its environmentally-dependent multiple developmental alternatives and facultatively heterotrophic and diazotrophic growth states. *N. punctiforme* lacks a flagellar apparatus that is present in the heterotrophic bacteria, although hormogonia are motile by a gliding mechanism. Thus, the 61 putative genes in the secretion and taxis category, while 3-fold less than those present in *P. aeruginosa*, may seem unusually high. *N. punctiforfme* has 3-5 copies each (21 total) of genes encoding homologs of the chemotaxis CheA, CheB and CheW signal transduction proteins, the CheD methyl-accepting protein and the CheR methylase, all which are likely to be involved in hormogonium tactic behavior. In addition, there are 11 genes encoding Sec, Sec-independent, and general protein secretion pathways.

In general, the genome of *N. punctiforme* reflects its complex life style. Only a few anticipated genes are absent in the unfinished database and they may be present in the completed sequence. Those genes involved in the phenotype defining energy, carbon and nitrogen metabolism categories will be discussed further in a subsequent section. Clusters of genes are present that were unanticipated based on the known *N. punctiforme* phenotype and there are notable multigene families; these will also be discussed later.

The state of the genome in comparisons to other cyanobacteria

A significant part of the information of a genome resides outside of genes. Protein binding sites may determine the regulation of gene transcription and the compact structure of the genome within the cell. The modification state of DNA may regulate the replication of the genome and aid in DNA repair. Transposable elements and repeated sequences may increase the plasticity of the genome and contribute to rapid changes over evolutionary time. In each of these regards, the genome of *N. punctiforme* is strikingly different from those of most bacteria.

Underrepresented sequences

One way to look for sequences of functional importance is to focus on those that are present in the genome significantly more or less frequently than would be expected by chance. The relative abundances of specific dinucleotides have been shown to be stable over related taxa (Karlin et al. 1997), and it is evident that the most underrepresented dinucleotides in the *N. punctiforme* genome (CG and TA) are also those of other cyanobacterial genomes (Table 4). It should be noted that the TA dinucleotide is underrepresented in most eubacterial genomes that have been examined (Karlin et al. 1997).

Table 5: Probable DNA modification enzymes of *N. punctiforme*

Recognition sequence	*Anabaena*[a]	*Synechocystis*[a]	prototype	Most similar protein[b]	
Solitary methyltransferases					
GATC	DmtA	MbpA	Dam, M.*Mbo*I	*Anabaena*	10^{-116}
GGCC	DmtB	Sll0729	M.*Hae*III	*Anabaena*	10^{-123}
CGATCG	DmtC	SynMI	M.*Pvu*I	*Anabaena*	$<10^{-200}$
rCCGGy	DmtD	-	M.*Cfr*10I	*Anabaena*	10^{-176}
Restriction/modification systems					
AGATCT	-	-	M.*Bgl*II	*Bacillus*	10^{-104}
			R.*Bgl*II	*Bacillus*	$2{\cdot}10^{-52}$
GrCGyC[c]	-	-	M.*Acy*I	*Hemophilus*	$3{\cdot}10^{-94}$
			R.*Acy*I	*Herpetosiphon*	$4{\cdot}10^{-67}$

[a]Protein (or predicted protein) from indicated cyanobacterium with strong similarity to proposed *N. punctiforme* modification enzyme.
[b]Organism with protein (or predicted protein) most similar to proposed *N. punctiforme* modification enzyme. BLAST **E** score is shown.
[c]The proposed restriction enzyme may be inactive, owing to stop codon near 5' end of region similar to corresponding restriction enzymes.

Palindromic hexanucleotide sequences are generally underrepresented in the genome of *N. punctiforme*, some extremely so. The most underrepresented hexanucleotides are sites for restriction enzymes that have been found in some species of *Nostoc* or *Anabaena*. This phenomenon, previously noted in the *E. coli* genome (Elhai 2001), points to a remarkable degree of DNA exchange amongst the Nostocaceae. Bias against palindromic sequences is seen also in tetranucleotides but not pentanucleotides. The only markedly underrepresented pentanucleotide is GGwCC, the recognition site for the common cyanobacterial restriction enzyme *Ava*II.

The foregoing is no less true for the genome of *Anabaena* PCC 7120; however, not all cyanobacteria have genomes biased against palindromic oligonucleotides. The genomes of *P. marinus* MED4 and *Synechococcus* WH8102 exhibit no bias for or against palindromic sequences. The genome of *Synechocystis* PCC 6803 lies midway in this regard between the extremes of *N. punctiforme* on one hand and the marine cyanobacteria on the other.

Restriction/modification systems and solitary DNA methyltransferases

Given the probable effect restriction systems have had on the makeup of the *N. punctiforme* genome, it is not surprising to find within the genome evidence for the transient residence of different DNA methyltransferases. *N. punctiforme* currently has one known and one suspected Type II restriction/modification system: those recognizing the same sites as *Bgl*II and *Acy*I (Table 5). The latter may be inactive. In addition, however, the genome has parts of five probably nonfunctional Type I restriction/modification systems and five Type II methyltransferases of unknown specificity and function. None of these systems show similarity to gene products predicted from the *Anabaena* PCC 7120 genomic sequence but often very high similarity to enzymes from other bacteria. These are prime examples of genes evidently acquired by horizontal transfer, preserved because they either conferred temporary selective advantage on their host (Bickle and Kruger 1993) or actively maintained their parasitic presence (Kobayashi et al. 1999).

In striking contrast to the lack of similarity between the restriction/modification proteins and known cyanobacterial proteins, four DNA methyltransferases without corresponding restriction enzymes (solitary methyltransferases) predicted from the *N. punctiforme* genomic sequence are nearly identical to those found in *Anabaena* PCC 7120 (Table 5; Matveyev et al. 2001). Three of these (DmtA, DmtB, and DmtC) appear to be widely distributed amongst cyanobacteria. The genome of *N. punctiforme* is thus almost certainly highly methylated, at sites common to most cyanobacteria (GATC, GGCC, CGATCG), another site common perhaps to the Nostocaceae (rCCGGy; methylated by DmtD, Matveyev et al. 2001), and sites peculiar to itself (AGATCG, GrCGyC, and perhaps several others from degrading Type I restriction/modification systems).

Multiple occurrences of HIP1 sequences

By far the most frequently occurring oligonucleotide is the octanucleotide sequence GCGATCGC, called HIP1 (Robinson et al. 2000). The sequence is widespread amongst cyanobacteria, but its function remains unknown. HIP1 occurs with a frequency of once every 1200 bp in the *N. punctiforme* genome, or about once every 800 bp if one counts frequently occurring sequences with at least 6 matches to the consensus sequence (Table 4). The distribution of HIP1 sites is nearly random in the genome, except for distances less than 40 bp, indicating that the selective pressure favoring HIP1 sites does not operate when a site already exists in close proximity. Although the function of HIP1 sites in cyanobacteria remains a mystery, it is interesting that internal to the sites are two methyltransferase targets (CGATCG and GATC) common to most, if not all, those cyanobacteria that carry HIP1 sites.

HIP1 sequences are found with a comparable frequency in *Anabaena* PCC 7120 and somewhat higher

Table 6. Ranked order of octanucleotide frequencies in cyanobacterial genomes[a]

	***N. punctiforme* ATCC29133**		***Anabaena* PCC 7120**		***Synechocystis* PCC 6803**	
	Octamer	Mean interval	Octamer	Mean interval	Octamer	Mean interval
1	**GCGATCGC**	1215	**GCGATCGC**	1267	**GGCGATCG**	911
2	**CGATCGCT**	2953	**AGCGATCG**	3209	**CGATCGCC**	923
3	**AGCGATCG**	2967	**CGATCGCT**	3301	**GCGATCGC**	1130
4	**CGATCGCA**	3051	**CGATCGCC**	3603	**GATCGCCA**	2430
5	**TGCGATCG**	3123	**GGCGATCG**	3653	**TGGCGATC**	2464
	AAAAATTA	7972	TAATTTTT	7911	ATCGCCAA	5134

	***Synechococcus* WH8102**		***P. marinus* MED4**	
	Octamer	Mean interval	Octamer	Mean interval
1	TGCTGCTG	3768	AAAAAAAT	2017
2	GCTGCTGG	4194	TTTTTTAA	2080
3	GCTGCTGC	4200	TTTTTAAA	2085
4	GCAGCAGC	4821	TTTAAAAA	2098
5	CAGCAGCA	4889	ATTTTTTT	2106

[a]The five most frequently occurring octanucleotides are shown for each genome, along with the mean distance in basepairs between occurrences. Sequences differing from the canonical HIP1 sequence by fewer than two nucleotides are shown in bold. Also shown is the most frequently occurring instance of an octanucleotide unrelated to the HIP1.

frequency in *Synechocystis* PCC 6803 but, surprisingly, almost not at all in the marine strains (Table 6). The consensus sequence in *Synechocystis* PCC 6803 is more relaxed in internal positions than in *N. punctiforme* and *Anabaena* PCC 7120, but it exhibits a strong preference for a flanking 5' G and 3'C not seen in the HIP1 sequences of the filamentous cyanobacteria.

Contiguous repeated sequences

Contiguous repeated sequences have been discovered serendipitously in some *Anabaena* PCC 7120 sequences (e.g. Mazel et al. 1990; Holland and Wolk 1990), but their prevalence in the genome of *N. punctiforme* and *Anabaena* PCC 7120 is astonishing. Such sequences, where the repeated region is at least 20 bp, account for about 1.5% of the total DNA within *N. punctiforme* (approximately 7.5% of the totla DNA in intergenic sequences). Many contiguously repeated sequences are found at multiple sites in the genome. The most frequently encountered sequences are shown in Table 7. All the most frequent unit sequences are heptamers, and they tend to fall into families, indicating that proteins recognizing them may have degenerate specificities. The sequences common in *N. punctiforme* overlap considerably with those found in *Anabaena* PCC 7120, but their frequencies are quite different (Table 7). Two of three previously named heptameric *s*hort *t*andemly *r*epeated *r*epetitive sequences, STRR1 and STRR2 (Mazel et al. 1990), occur frequently in the *N. punctiforme* genome, but STRR3 does not appear at all, nor does the 37-bp repeat (LTRR) noted in *Anabaena* PCC 7120 (Masepohl et al. 1996). We have named four repeated sequences STRR4, STRR5, STRR6, and STRR7, one of which (STRR5; Angeloni and Potts 1994) has been previously noted in a strain of *Nostoc*.

The heterocyst-forming cyanobacteria stand apart from other cyanobacteria in the prevalence and type of contiguous repeated sequences. *Synechocystis* PCC 6803 and *Synechococcus* WH8102 have few repeated sequences except those based on triplets or their multiples. These presumably are repeated codons. *P. marinus* has many nontriplet repeats, but almost all are extremely AT-rich and may merely reflect the presence of AT-rich regions in the genome. *E. coli* is similarly lacking in contiguous repeated sequences, and it has been suggested that prokaryotes generally have few such repeats (Field and Wills 1998). The excess of heptameric repeats thus appears to be unique to the heterocyst-forming cyanobacteria. No clue has yet been provided as to their functions.

Insertion sequences

The genomic sequence of *N. punctiforme* in its current state carries well over 150 ORFs that are significantly similar, in whole or in part, to transposases. It is difficult to arrive at an accurate count, in part because contigs often end within transposases, owing to the difficulty of assembling these ends. Furthermore, the transposases exist in multiple states of degradation. It is clearly evident, however, that the genome has suffered waves of infection by foreign sequences that insert themselves multiple times in the genome and progressively mutate to nonfunctionality and eventual oblivion. Early on, copies of the gene encoding the transposase suffer mutation, but the insertion sequence of which it is a part continues to transpose, using transposase encoded by sister sequences. Once the last transposase gene is rendered nonfunctional, the insertion sequence is dead, a target of mutation and insertion by other insertion sequences. All stages in this progression of events are evident in the genomic sequence. The functionality of an insertion sequence can be

Table 7: Most frequently occurring contiguously repeated sequences in *N. punctiforme*

		Sites in *N. punctiforme* ATCC 29132[c]		Related sites in *N. punctiforme*[d]		Sites in *Anabaena* PCC 7120[c]		
Rank[a]	Unit sequence[b]	Sites in genome	Avg # of iterations	Sites in genome	Avg # of iterations	Sites in genome	Avg # of iterations	Rank[a]
	AATGACh (STRR2)							
1	AATGACA	69	5.0	242	5.0	10	4.7	35
2	AATGACT	63	5.1	237	4.9	33	5.8	8
5	AATGACC	39	5.0	236	4.9	14	4.2	26
	AATTCCC (STRR4)							
4	AATTCCC	41	4.7	190	4.7	5	5.4	75
7	AATGCCC	37	4.5	179	4.8	0	-	-
3	AATTACG (STRR5)	45	4.2	98	4.3	8	4.5	43
	AdTCCCC (STRR1)							
6	ATTCCCC	39	4.7	180	4.6	8	5.0	48
21	AATCCCC	19	5.1	183	4.7	58	6.1	2
28	AGTCCCC	15	4.7	9.7	4.6	54	5.3	3
8	AGCAGGGG (STRR6)	29	4.4	50	4.2	5	3.8	78
30	AAAATTC (STTR7)	13	3.7	128	4.1	75	4.1	1

[a]Rank of unit sequence ordered by the number of sites in the genome of *N. punctiforme*. The eight highest ranked unit sequences are shown, plus the three highest ranked of *Anabaena* PCC 7120.

[b]Predominant unit sequence of contiguous repeat. The unit may be considered to be any of the possible circular permutations of the sequence and its inverse. The permutation appearing lowest in the alphabet was preferred to facilitate matching. STRR1 and STRR2 are according to Mazel et al. (1990).

[c]Sites in the genome where there appears a contiguous repeat with the given unit sequence predominating.

[d]Sites in the genome where there appears a contiguous repeat where the predominating unit sequence is no more than one base removed from the given unit sequence.

discerned from the sequence only by the presence of multiple copies of identical sequences or nearly identical sequences with identical termini. By this conservative criterion, *N. punctiforme* has close to six active insertion sequences, each with multiple insertions.

Almost all of the transposases found encoded in the *N. punctiforme* genome are most similar to proteins reported in other cyanobacteria (often *Synechocystis* PCC 6803), and of these, roughly two-thirds are most similar to proteins from *Anabaena* PCC 7120. However, there are exceptions. For example, one predicted *N. punctiforme* protein is most similar to a Tn*21*-like transposase from *Bacillus anthracis*, with a BLAST score of less than 10^{-200} and no significant similarity to any protein in *Anabaena* PCC 7120 or *Synechocystis* PCC 6803. Neither *P. marinus* MED4 or *Synechococcus* WH8102 have any open reading frames recognizable as transposases or insertion sequence proteins.

Genetic information specific to energy, carbon and nitrogen metabolism.

Energy and carbon metabolism

Photosynthetic electron transport. In *N. punctiforme*, as in other cyanobacteria, the genes encoding proteins of reaction center complexes, the cytochrome b_6/f complex and electron carriers between the complexes are not clustered in the genome, although there appears to be some operon structure. This is in contrast to large gene clusters in the purple, non-sulfur, anoxygenic photosynthetic bacteria, such as *Rhodobacter sphaeroides* (Naylor et al. 1999). Except for *psbA* (32 kDa or D1 protein) and *psbD* (34 kDa or D2 protein), there are single copies of genes encoding PS2 reaction center proteins. There are four complete and three truncated copies of *psbA* and one complete and one truncated copy of *psbD*. The truncated copies appear at the ends of contigs and may be found to be complete genes after the sequencing of the genome is complete. These two genes are well known members of multigene families in cyanobacteria and differential transcription of *psbA* and *psbD* in response to varying light intensities has been documented (Golden 1994). The genes encoding PS1 reaction center proteins occur as single copies with *psaA* and *psaB* collocated in the genome in a putative operon. Two additional truncated copies of *psaB* also appear at the ends of their respective contigs. Genes encoding proteins of the cytochrome b_6/f complex appear as pairs in putative operons with the following linkages: *petB* (cytochrome *b*)-*petD* (subunit IV), *petC* (Reiske Fe/S protein)-*petA* (cytochrome *f*), and *petE* (plastocyanin)-*petJ* (cytochrome *c*553); there is a second solitary copy of *petJ*.

In the presence of sufficient copper, cyanobacteria synthesize plastocyanin, but when starved for copper, cytochrome *c*553 is synthesized as the soluble electron carrier between the cytochrome b_6/f and PS1 complexes (Morand et al. 1994). There is no evidence for linkage of the gene pairs. *N. punctiforme* contains a single gene (*petH*) encoding ferredoxin-NADP oxidoreductase and 15 putative ferredoxin genes that include 9 encoding 2Fe-2S *petF* type and 6 encoding a 4Fe-4S type.

Pigment synthesis. In cyanobacteria, tetrapyrrole biosynthesis is initiated with glutamyl-tRNA as the substrate for δ–aminolevulinic acid synthesis (Beale 1999). There are four genes encoding glutamyl-tRNA synthetase and single genes encoding glutamyl-tRNA reductase and glutamate 1-semialdehyde aminostransferase in the *N. punctiforme* genome, consistent with this biosynthetic pathway. It is of interest that genes encoding oxygen-dependent and oxygen-independent coproporphyrinogen III oxidases are also present in the *N. punctiforme* genome, indicating the potential for porphyrin synthesis in both oxic and anoxic environments. *N. punctiforme* contains genes encoding proteins involved in light-dependent regulation of PE synthesis, based on similarity to genes identified in *Fremyella diplosiphon* (also known as *Calothrix* sp. strain PCC 7601) (J. Cobley, personal communication).

N. punctiforme exhibits Type II CCA, but the question remains whether it also has the potential for Type III CCA. Type III CCA is associated with two or more clusters of *cpc* genes encoding PC. Unfortunately, the sequence of the *cpc* genes is incomplete in the current *N. punctiforme* database. The *cpcCDEFGH* genes encoding linker proteins are present in single copies, but of the genes encoding the apoproteins for chromophore binding, the *cpcB* gene is missing and only a fragment of *cpcA* at the end of a contig can be unequivocally identified. Therefore, we cannot predict the number of *cpcBA* operons. Two clusters of *cpeBA* genes encoding the PE apoprotein subunits can be detected, but this may be an assembly error. Insertions into *cpeBA* result in a *N. punctiforme* mutant lacking detectable PE, implying a single functional copy (F.C. Wong, E.L. Campbell and J.C. Meeks, unpublished). It is not clear why automated cloning and sequencing of genes involved in phycobiliprotein synthesis has proven difficult in *N. punctiforme*. Type III CCA is also characterized by a signal transduction pathway involving RcaE, RcaF and RcaC (Bhaya et al. 2000). ORFs with regions of similarity to the corresponding structural genes are present in the *N. punctiforme* genome, but the domain organization is often inconsistent. Thus, the presence of genes associated with Type III CCA is unresolved.

Respiratory electron transport. In cyanobacteria, the cytochrome b_6/f complex is shared by both respiratory and photosynthetic electron transport systems (Schmetterer 1994). Genes are present that are consistent with synthesis of both the multiprotein complex mitochondrial type-1 and the 1- to 2-subunit bacterial type-2 FAD-containing NADH dehydrogenases (NADH: plastoquinol oxidoreductase). A type-1 *ndhCKJ* contiguous cluster can be detected, similar to other cyanobacteria (Schmetterer 1994), but the other genes are unlinked. Four gene copies encoding the large subunit type-2 dehydrogenase are present. There appear to be at least 4 and possibly 6 copies of a *ctaCDE* operon encoding subunit proteins II, I and III, respectively, for an aa_3-type cytochrome *c* oxidase; *ctaD* is truncated 4 times, twice in each orientation, at the ends of contigs.

Carbon metabolism - assimilation. Carbon dioxide assimilation and hexose catabolism occur through the reductive (Calvin-Benson-Bassham pathway) and oxidative pentose phosphate pathways in cyanobacteria (Smith 1982). The enzymes specific for the reductive pathway are phosphoribulokinase (PRK) and ribulose 1,5-bisphosphate carboxylase/oxygenase (rubisco). The genes encoding both enzymes are present in single copies. The rubisco genes are clustered with the following transcriptional orientation: *rbcL-rbcX-rbcS-orfH1-orfH2-orfH3-rca*, where L is the rubisco large subunit, X is a protein with ambiguous function, S is the rubisco small subunit, *rca* is rubisco activase and *orfH1-3* encode hypothetical proteins with similarity in the *Synechocystis* PCC 6803 genome. A putative rubisco transcriptional regulator (*rbcR* or *cbbR*) is located near a cluster of 7 genes encoding proteins involved with a carbon dioxide concentrating mechanism (Ccm) in the following orientation: *ccmK3-ccmK2-ccmL-ccmM-ccmN-fpg-ccmK1-ccmML-rbcR*, where *fpg* is a putative formamidopyrimidine-DNA glycosylase and *ccmML* is found between *ccmM* and *ccmL* in Proteobaceria. Genes encoding *ccmK4* and *ccmK5* are located elsewhere. *N. punctiforme* has at least 5 copies of carbonic anhydrase, an essential Ccm enzyme, located in solitary positions throughout the genome. It is not clear if a specific bicarbonate transport system is present as part of the Ccm because BLAST result candidate ORFs show a high degree of similarity to both identified bicarbonate and nitrate/nitrite transporters. Mutation of *N. punctiforme* and phenotypic characterization will be required to unequivocally establish specificity.

Carbon metabolism - catabolism. The enzymes specific for the oxidative pentose phosphate pathway are glucose-6-phosphate dehydrogenase (G6PD) and 6-phosphogluconate dehydrogenase (6PGD). The *N. punctiforme* genome has three copies each of genes encoding these proteins and this redundancy is thus far unprecedented in cyanobacteria. One copy of the gene encoding G6PD (*zwf*) lies within an operon (*opc*; Summers et al. 1995) along with three other gene in the following orientation: *fbp-tal-zwf-opcA*. The gene *fbp* encodes fructose1,6-bisphosphatase, *tal* encodes transaldolase and *opcA* encodes a protein allosteric effector of G6PD (Hagen and Meeks 2001). The organization of these genes is the

same in *Anabaena* PCC 7120, but they are organized differently in other cyanobacteria. *Synechococcus* WH8102 and *P. minor* MED4 have an apparent *zwf-opcA* operon with the other genes localized elsewhere. In *Synechococcus elogatus* PCC 7942, *tal* is located apart from the cluster (Newman et al. 1995) and in *Synechocystis* PCC 6803 the four genes are all unlinked in the chromosome. In *N. punctiforme*, one *gnd* encoding 6PGD is located elsewhere, while the other two copies are collocated with the two additional copies of *zwf*. The two additional copies of G6PD share 67% similarity to each other, but only 49-53% similarity to the *opc* operon *zwf*-encoded G6PD. The two 6PGD proteins encoded by genes linked to *zwf* show 78% similarity to each other and about 45% similarity to the solitary 6GPD. Expression studies have not been done to identify the *gnd* that is predominantly transcribed. Since inactivation of *zwf* in the *opc* operon yields defects in nitrogen fixation and dark heterotrophic growth of *N. punctiforme*, the role of the additional copies of *zwf* and *gnd* is unclear.

Nitrogen assimilation

Nitrogenase genes. Nitrogen fixation is mediated by the nitrogenase enzyme complex that requires on the order of 20 gene products for synthesis and assembly (Dean and Jacobson 1992). The structural genes for dinitrogenase (*nifD* and *nifK*) and dinitrogenase reductase (*nifH*) are well conserved among all bacteria, including cyanobacteria, where these genes form an operon. The organization and order of a large cluster of *nif* and *nif*-related genes in cyanobacteria is highly conserved. (Thiel et al. 1997, 1998) In *N. punctiforme*, *Anabaena* PCC 7120, *A. variabilis* ATCC 29413 and probably *Synechococcus* RF-1, the gene order is *nifB-fdxN-nifS-nifU-nifH-nifD-nifK-orf-nifE-nifN-nifX-orf-orf-nifW-hesA-hesB-fdxH*. (Buikema and Haselkorn 1993; Thiel et al. 1997, 1998; Huang et al. 1999).

The *N. punctiforme nifD* gene has a 24-kb excision element near the 3' end. The *nifD* gene of *Anabaena* PCC 7120 is interrupted in exactly the same location by an 11-kb excision element (Golden et al. 1985) and the *nifD* gene in *A. variabilis* also has an 11-kb excision element (Brusca et al. 1989). The 11-kb and the 24-kb elements share a highly conserved excisase gene (*xisA*) located at the beginning of the element that is required for excision of the element during heterocyst differentiation, and a small ORF of unknown function. Otherwise, there is no similarity between the two *nifD* elements. The *N. punctiforme* element contains homologs of two genes of unknown function in *Synechocystis* PCC 6803 and a homolog of a gene of *Anabaena* sp. called protein X (Sato 1994). The other 14 ORFs in the 24-kb element have no similarity to cyanobacterial genes; however, one ORF has similarity to bacterial reverse transcriptase genes (*ret*), supporting the suggestion that such elements may be remnants of lysogenic phage (Ramaswamy et al. 1997).

The *nif* region of *N. punctiforme* differs in several other respects from that of *Anabaena* PCC 7120. *N. punctiforme* lacks the 55-kb excision element in the *fdxN* gene that is present in *Anabaena* PCC 7120 (Golden et al. 1988) Upstream of *nifH* in *N. commune* (Potts et al. 1992) and in *N. punctiforme* there is a hemoglobin-like gene called cyanoglobin (*glbN*) whose function is not known. Phylogenetic analysis of the *nifH* and *nifD* genes of *N. punctiforme* indicates that they are most closely related to their homologs in *N. commune* (T. Thiel, unpublished). About 10 kb upstream of *nifB* in *N. punctiforme* are homologs of *nifP*, *nifZ* and *nifT*. Just downstream of the major *nif* cluster are genes for an uptake hydrogenase, including *hupS* and *hupL*. In *Anabaena* PCC 7120, the *hupL* gene is distant from the *nif* region and is interrupted by a 10.5 kb excision element (Carrasco et al. 1995). This element is absent in *hupL* of *N. punctiforme*, as well as several other strains of cyanobacteria (Tamagnini et al. 2000).

The *nif* genes described above encode the molybdenum-dependent nitrogenase that functions exclusively in heterocysts. *A. variabilis* and a few closely related cyanobacterial strains (but not *Anabaena* PCC 7120) have two alternative nitrogenases (Thiel 1998; Thiel and Pratte 2001). One is a vanadium-dependent nitrogenase that functions only in the absence of Mo under conditions in which heterocysts form (Thiel 1996). The other is a Mo-dependent nitrogenase that functions in vegetative cells under strictly anoxic conditions (Thiel et al. 1997). *N. punctiforme* has only one complete set of *nif* genes and those appear to encode the heterocyst specific Mo-nitrogenase. *N. punctiforme* has two additional copies of *nifH* and one additional copy of *nifE* and *nifN*. One of the copies of *nifH* is immediately upstream of *nifE* and *nifN* forming a contiguous cluster of three genes. The third copy of *nifH* has no other *nif* genes nearby. It is not unusual for bacteria to have multiple copies of *nifH*: *Anabaena* PCC 7120 has two genes and *A. variabilis* has four. Phylogenetic analysis indicates that the second copy of *nifH* in *N. punctiforme* (near the second copy of *nifEN*) clusters with the *N. punctiforme* and *N. commune nifH* genes that are part of the major *nif* cluster; however it is less like those two genes than they are like each other (T. Thiel, unpublished). The third *nifH* in *N. punctiforme* is closely related to the *nifH* gene in *A. variabilis* that appears to encode the dinitrogenase reductase of the V-nitrogenase (T. Thiel, unpublished). None of the *nifH* genes in *N. punctiforme* is closely related to the second copy of *nifH* in *Anabaena* PCC 7120.

Nitrate and nitrate utilization. In *Anabaena* PCC 7120 the genes involved in the uptake and reduction of nitrate and nitrate comprise a cluster: *nirA* (nitrite reductase)-*nrtA-nrtB-nrtC-nrtD* (ABC transporter)-*narB* (nitrate reductase) (Frías et al. 1997; Cai and Wolk 1997). In *N. punctiforme* the *nirA* and *narB* genes are separated by a single gene that appears to encode a permease that transports nitrate/nitrite. This permease is similar to the nitrate/nitrite transporter

gene, *nrtP*, identified in the marine strain, *Synechococcus* sp. strain PCC 7002, (Sakamoto et al. 1999), and to another gene in that same family (*napA*) found in marine strains *Trichodesmium* sp. and *Synechococcus* sp. strain WH7803 (Wang et al. 2000). A gene for this type of permease is not present in the genomes of *Anabaena* PCC 7120, *Synechocystis* PCC 6803, or *P. marinus* MED4 (the latter of which cannot grow on nitrate or nitrite). The presence of the permease in *N. punctiforme* suggests that this type of nitrate/nitrite transporter is not necessarily associated with marine strains of cyanobacteria. In addition to the *nrtP*/*napA* permease in *N. punctiforme*, there is a cluster of 4 genes on 3 short contigs that together comprise an ABC transporter with over 90% amino acid similarity to the *nrtABCD* nitrate transporter of *Anabaena* PCC 7120. Thus, it appears likely that *N. punctiforme* has two independent nitrate/nitrite transport systems.

Ammonium assimilation. Three genes putatively encoding distinct ammonium transport systems are present in the *N. punctiforme* genome. The enzymes glutamine synthetase (GS) and glutamate synthase (GOGAT) are essential for the assimilation of exogenous and nitrate- or dinitrogen-derived ammonium in cyanobacteria (Flores and Herrero 1994). There are two genes in *N. punctiforme* with similarity to GS: one shows 76% and over 90% amino acid identity to *glnA* of *Synechocystis* PCC 6803 and *Anabaena* PCC 7120, respectively. The other gene is weakly similar to *glnA*, but has greater similarity (46% identity) to *tdnQ* in *Pseudomonas putida*. The *tdnQ* gene encodes a putative amino group transferase that that is thought to be involved in the pathway for conversion of aniline to catechol in *P. putida* (Fukumori and Saint 1997). A ferredoxin-dependent GOGAT with over 90% amino acid similarity to GOGAT in *Anabaena* PCC 7120 is present in *N. punctiforme*. There are no genes for a second NADH dependent GOGAT in either *N. punctiforme* or *Anabaena* PCC 7120 (Martin-Figueroa et al. 2000) as has been described for *Plectonema* (Okuhara et al. 1999). In addition to GS/GOGAT, ammonium could be assimilated directly into glutamate via glutamate dehydrogenase (*gdhA*). GdhA of *N. punctiforme* shows moderate amino acid similarity (63%) to GdhA of *Synechocystis* PCC 6803.

Genes involved in heterocyst differentiation. Genes identified in *Anabaena* PCC 7120 as important for heterocyst differentiation are also present in *N. punctiforme*. Many of these genes encode proteins with over 90% amino acid sequence identity between the two strains. Among this group are *ntcA* (a regulatory protein that is essential for many aspects of nitrogen metabolism) (Frías et al. 1994), *hanA* (encoding HU, a histone-like protein) (Khudyakov and Wolk 1996), *devH* (a putative DNA-binding protein) (Hebbar and Curtis 2000), *hetR* (a protease that regulates an early step in heterocyst differentiation) (Zhou et al. 1998), *patB* (affects pattern formation) (Liang et al. 1993), *devR* (heterocyst maturation) (Campbell et al. 1996) and *devBCA* (an ABC transporter required for heterocyst envelope formation protein) (Fiedler et al. 1998). There is a second copy of *devA* (90% amino acid identity), but not *devBC*.

Several genes encode proteins with about 60-70% amino acid identity (70-85% similarity) between the strains. These include *hetF* (a positive regulator of heterocyst differentiation) (Wong and Meeks 2001), three genes involved in heterocyst envelope synthesis, *hepA* (Holland and Wolk 1990), *hepK* (a sensor histidine kinase of a two-component regulatory system) (Zhu et al. 1998) and *hglK* (Black et al. 1995), as well as *patA* (affects pattern formation) (Liang et al. 1992), *hetM* (polyketide synthase) and *hetI* (unknown function). *hetM*, *hetN* and *hetI* are part of a contiguous gene cluster in *Anabaena* PCC 7120 (Black and Wolk 1994). While the homologs of *hetM* and *hetI* are present in a cluster in *N. punctiforme*, the most similar homolog of *hetN* (a ketoacyl reductase) is not in this cluster, but is present on a different contig. This gene, encoding a protein with 70% amino acid similarity to *hetN*, does not have homologs of *hetM* or *hetI* nearby. Interestingly, however, there is a gene with similarity to a ketoacyl reductase from *Mycobacterium tuberculosis* (54% amino acid similarity) located between *hetM* and *hetI* in *N. punctiforme*. Another polyketide synthase with about 50% amino acid similarity to *hetM* is present on a different contig.

A few genes involved in heterocyst differentiation show relatively weak similarity to their counterparts in *Anabaena* PCC 7120 including *hepC* (required for heterocyst envelope synthesis) (Zhu et al. 1998), and *hetP* and *hetC* (similar to ABC protein exporters and required early in heterocyst differentiation) (Khudyakov and Wolk 1997). The latter two genes are contiguous in *Anabaena* PCC 7120 but not in *N. punctiforme*. The *hetP* homologs in the two strains have about 70% amino acid similarity; however, there are gaps in the alignment. There are two copies of *hetC*-like genes in *N. punctiforme* with different sequences but both are about 66% similar (amino acids) to *hetC* in *Anabaena* PCC 7120. One possible explanation for the weak similarity of some of these genes between the two strains is that they originally served different functions in the two strains and evolved to take on the functions required for heterocyst differentiation.

The *patS* gene that is thought to produce a small peptide inhibitor of heterocyst formation (Yoon and Golden 1998) encodes an ORF of only 13 amino acids in *N. punctiforme*. In *Anabaena* PCC 7120, *patS* encodes two possible ORFs of 13 or 17 amino acids. Mutation of either or both met codons in *Anabaena* PCC 7120 did not prevent normal heterocyst suppression; however, the mutant *patS* genes were under control of the very strong *glnA* promoter on a plasmid, perhaps allowing expression from a cryptic start site (J. Golden, personal communication). The presence of only the 13 amino acid ORF in *N. punctiforme* and the good ribosome binding site just upstream (which is also present in *Anabaena* PCC

7120) indicates that the precursor peptide is likely 13 amino acids.

Other notable characteristics of the *N. punctiforme* genome.

Desiccation response.
Many terrestrial cyanobacteria show a resistance to water deficit. One mechanism that contributes to desiccation tolerance is the synthesis of non-reducing disaccharides such as trehalose and sucrose. *N. punctiforme* lacks a homolog of trehalose-6-phosphate synthase (*otsA*). Only plants and cyanobacteria can synthesize sucrose and their sucrose-6-phosphate synthases (SpsA) are highly conserved. Homology searches revealed at least 20 proteins that show similarity to sucrose-6-phosphate synthase, and a *spsA* homolog has tentatively been identified (L. Curatti and G. Salerno, personal communication). A *spsA* homolog is present in desiccation-tolerant *Nostoc commune* DRH1. *N. punctiforme* also contain genes putatively encoding invertase and the bifunctional sucrose synthase, which is consistent with a capacity to synthesize and degrade sucrose. *Synechocystis* PCC 6803, which has the capacity to withstand air-drying, contains *otsA* and *ggpS* (involved in synthesis of the compatible solute glucosyl-glycerol), *spsA* and sucrose 6-phosphate phosphohydrolase (*spp)*. *Synechocystis* PCC 6803 also contains four other genes involved in glucosyl-glycerol metabolism; only one of these (slr0530) has a homolog in *N. punctiforme*. The genome of *N. punctiforme* also lacks any homolog of the water stress protein gene *wsp* which is specific to the form species *N. commune* (Wright et al. 2001).

Other genes important in desiccated cells include those for superoxide dismutase, catalase, and DNA repair enzymes. *N. punctiforme* contains three *sodA*-like genes as well as a homolog of *N. commune sodF* that represents the third most abundant protein in the latter strain. Catalase (*katG*) in *Synechocystsis* PCC 6803 shares strong homology with counterparts in a range of other bacterial species and a gene with weak similarity is present *N. punctiforme*. *N. punctiforme* contains several examples of hydrophilins such as a HSP70-class of molecular chaperone (rod-shape protein) potentially involved in cell-wall biogenesis. Hydrophilins are characterized by high glycine content (>6%) and a high hydrophilicity index (>1.0). The criterion that defines hydrophilins seems to be an excellent predictor of responsiveness to hyperosmosis (Garay-Arroyo et al. 2000). Thus, consistent with its phenotype of tolerating slow drying, *N. punctiforme* may not have the capacity to tolerate rapidly alternating desiccation cycles that is characteristic of many terrestrial cyanobacteria.

Circadian rhythms.
Cyanobacteria are now known to exhibit circadian rhythms (Golden et al. 1998). A gene cluster encoding proteins denoted KaiA, KaiB and KaiC is essential for the rhythm (Ishiura et al. 1998); KaiC has similarity to the RecA superfamily ATPases. A complex sensor histidine kinase (CikA) with an attached chromophore is a component of the environmental sensing pathway involved in entrainment (Schmitz et al. 2000). The *N. punctiforme* genome contains homologs of *kaiABC* in a cluster, plus *cikA* elsewhere, although there is no physiological evidence of a circadian rhythm.

Multigene families.
The multigene families of *N. punctiforme* can be organized into 5 broad groupings: environmental sense and response, transcriptional regulation, transport, transposition (previously discussed), and those that encode large complex multidomain proteins.

Environmental sense and response. Environmental sensing and response in bacteria occurs primarily through protein histidine-aspartate phosphorelay systems generally referred to as two-component regulatory systems (Hoch and Silhavy 1995). The signal sensing and transmitting component consists of a sensor histidine kinase (or transmitter) which autophosphorylates an invariant histidine residue in an ATP-dependent mechanism in response to an environmental signal. The phosphorylated transmitter transfers the phosphate to an invariant aspartate residue in a cognate receiver protein called a response regulator (or receiver). The response regulators most often have an output domain that defines a DNA-binding motif through which the protein regulates transcription, although some have no output domain. There also exists a class of complex signal transducers that contains both transmitter and receiver domains.

N. punctiforme has an unusually high number (255) of combined two-component signal transduction proteins. By comparison, *Synechocystis* PCC 6803 has 42 transmitters and 38 response regulators and *B. subtilis* has 38 transmitters and 34 response regulators, for totals of 80 and 72, respectively. *E. coli* contains genes encoding 23 simple transmitter proteins, 32 response regulator proteins and 5 complex transmitter-response regulator proteins; a total of 62. In *E. coli*, the cognate sensor kinase-response regulator for a specific environmental signal also tend to lie contiguously on the chromosome. The vast majority (88%) of the 153 *N. punctiforme* transmitter genes are unlinked to a response regulator. Single domain sensor histidine kinases constitute only about 53% of this class of genes, while genes encoding complex proteins consisting of transmitter-receiver, transmitter-receiver-transmitter, and transmitter-receiver-receiver-transmitter (and other combinations) constitute the remaining 47%. The 102 response regulator genes consist of 36% encoding receivers with no apparent output domain, while the remainder are characterized by a helix-turn-helix DNA binding output domain. The unusually high frequency of response regulators lacking output domains may reflect extensive operation of multiprotein phosphorelay signaling

systems, similar to those that control sporulation in *B. subtilis* (Ireton et al. 1993). The advantage of extended phosphorelay systems is integration of multiple environmental signals in the signaling pathway.

Included in the above analyses were the simple chemotaxis sensor histidine kinases and complex chromophore-binding sensor histidine kinase proteins with homology to CAA (*rcaE*) and circadian rhythm (*cikA*) responses. Additional complex chromophore-binding sensor histidine kinases are those that encode phytochrome homologs. *N. punctiforme* contains at least 6 genes encoding cyanobacterial phytochrome proteins (Yeh et al. 1997), 2 *cph1* and 4 *cph2*, plus 15 other phytochrome-like proteins; only four of the total phytochrome-like proteins lack obvious histidine kinase domains (J. C. Lagaris, personal communication). In addition to simple and complex sensor histidine kinases, *N. punctiforme* also contains 55 ORFs encoding eukaryotic serine/threonine protein kinases. The majority (62%) are single domain Ser/Thr kinsases, but 21 of these ORFs encode putative proteins with both Ser/Thr and His kinase domains. Their physiological roles remain undefined. It is likely that all of these protein kinases constitute a phylogenetic family of sensory transduction proteins.

In addition, the *N. punctiforme* genome contains at least 7 genes encoding adenylate cyclase involved in cAMP synthesis, perhaps in response to metabolite sensing.

As judged by the number of genes apparently encoding signal transduction proteins, the potential capacity of *N. punctiforme* to sense and respond to environmental changes is extraordinary; more so than any bacterium characterized to date. The nature and extent of the environmental signals remains to be determined. However, we speculate that much of the signal transduction capacity will be involved in the multiple developmental alternatives that *N. punctiforme* can express in response to environmental signals. The requirement in heterocyst development for DevR, a response regulator lacking an output domain (Campbell et al. 1996), implies that one or more multiprotein phosphorelay system is involved in that differentiation event.

Transcription. For quite some time, cyanobacteria were thought to be limited in the extent of transcriptional regulation in response to environmental changes. The presence of response regulators with output domains and the documented differential gene expression in heterocyst differentiation, nitrogen assimilation and CCA clearly negates that long held assumption (Tandeau de Marsac and Houmard 1993). Nevertheless, transcriptional regulation is poorly described in cyanobacteria (Curtis and Martin 1994). In bacteria, transcriptional regulation can be dictated by promoter sequence recognition by the σ subunit of DNA-dependent RNA polymerase and/or by the presence of transcriptional inhibitors or activators that are not a part of the RNA polymerase holoenzyme. The *N. punctiforme* genome contains 13 apparent alternative σ^{70} subunits in addition to the primary σ^{70} subunit. Only one alternative σ^{70} has been mutated in *N. punctiforme* leading to a discernable phenotype involving symbiotic interaction (Campbell et al. 1998). σ^{54} and the elements of the signal transduction pathway governing the expression in Proteobacteria of nitrogen responsive genes, including *nif* (Merrick and Edwards 1995), is absent in the *N. punctiforme* genome, except for *glnB* encoding the P_{II} protein.

The *N. punctiforme* genome contains at least 57 genes encoding ancillary transcriptional regulatory proteins in addition to the putative response regulators with output domains; 67% are classified only as predicted and have similarity to a variety of known regulatory proteins such as TetR, XylR and ArsA. There are 6 copies of genes with high similarity to those encoding AraC, 8 of LysR and 2 of MocR. These collective data imply that *N. punctiforme* has a substantial capacity for differential gene expression in response to a variety of environmental signals.

Transport. *N. punctiforme* has 262 ORFs encoding proteins that play an assigned role in transport of small organic and inorganic molecules across the cell membrane. There are 89 ORFs in the *N. punctiforme* genome that have been provisionally identified as encoding the ATPase domain of assigned and unassigned membrane-associated ATP-binding cassette transport systems (ABC transporters). In addition, there are 48 organic carbon and ion transporting permeases not associated with ABC transporters. There appear to be no representatives of the phosphotransferase system of enteric bacteria in the *N. punctiforme* genome.

There are two complete ATP-dependent phosphate transport systems, each comprising *pstS*, *pstC*, *pstA* and *pstB*. In addition, there is a contiguous cluster of genes, probably also involved in phosphate transport, with genes similar to *pstC* and *pstA* from *Archaeoglobus*, and to *pstB2* from *Synechocystis* PCC 6803. Also associated with this cluster is a gene with some similarity to *spbA* (encoding the periplasmic binding protein for sulfate transport) and another that is similar to *sphX* in *Synechocystis*, a gene that is regulated by SphR, which responds to phosphate limitation. There is a single sulfate transport system with two divergent copies of *sbpA* followed by *cysT* and *cysW*. The ATP-binding protein of the sulfate transport system, *cysA*, is on a different contig. The putative nitrate transport system is described in the section on nitrate and nitrate utilization. Three genes have similarity to a putative glutamine transporter. Molybdenum is probably transported by the products of genes that are similar to *modA* (encoding the periplasmic binding protein) and a fused ModBC protein that combines the function of the permease and the ATP-binding protein. Another ABC transporter has similarity to putative zinc and manganese transporters in other bacteria. An unusual ABC transporter comprising four genes has about 70% amino acid similarity to the *ptxABCD* genes that are thought to

function in phosphite transport in *Pseudomonas stutzeri* (Metcalf and Wolfe 1998). There are genes with similarity to sugar transport systems, particularly ribose and hexose transport and to peptide transport systems. Although there are many other genes with similarity to various components of ABC transporters, most of these are not associated with a complete set of genes known to be required for transport in other bacteria and, hence, may function in conjunction with other transport systems.

Multidomain proteins putatively synthesizing cyclic peptide toxins. *N. punctiforme* contains 62 ORFs encoding proteins involved in the apparent synthesis of cyanobacterial secondary products classified as microcystins. Microcystins are hepatotoxins that inhibit eukaryotic protein phosphatase activity; they are synthesized and released by a variety of unicellular and filamentous cyanobacteria (Dow and Swoboda 2000). Structurally, microcystins are hybrid cyclic peptide-polyketide molecules of molecular mass between 820 and 1044 Da. Microcystins are synthesized by the sequential activity of non-ribosomal peptide synthetases (NRPS) and polyketide synthases (PKS), together with chain or side chain modifying activities. The NRPS and PKS activities may be confined to respective single proteins or be collocated on multidomain proteins. The 34 genes encoding the multidomain hydrid NRPS-PKS proteins in *N. punctiforme* were erroneously identified by automated annotation as Acyl-CoA synthetase (AMP forming)/AMP-(fatty) acid ligases I. The ORFs range in size from 1,031 to 14,048 bp and 53 complex and simple ORFs are clustered in 2 sets of 3 genes and 1 set each of 4, 5, 6, 12 and 14 genes. The latter two sets at 46.78 and 49.22 kb constitute the largest common gene clusters in the *N. punctiforme* genome and come the closest to the definition of gene islands.

There is no precedence for the production of microcystins by the terrestrial *N. punctiforme*, so the detection of these genes was unanticipated. Essentially all production of cyanotoxins has been recorded in aquatic habitats, especially those experiencing dense growth blooms of cyanobacteria (Dow and Swoboda 2000). Marine cyanobacterial strains produce similar hybrid molecules; many are also halogenated and have biological activity (Sitachitta et al. 2000). The extent of production of such compounds may be more extensive in cyanobacteria than has been anticipated by culture and habitat sampling. Similar to antibiotic production by fungi and Gram positive bacteria, the physiological role and selective advantage for secondary product production by the individual organisms is uncertain.

Conclusions

The genome of *N. punctiforme* has many of the characteristics one would expect of a sequence that is highly plastic and in a state of flux. The genome has a conspicuous number of elements -- insertion sequences and multilocus repeats -- that can participate in genome rearrangements, duplications, and deletions. The inventory of transposases and DNA modification enzymes, and the paucity of restriction sites, indicate extensive exchange of DNA amongst cyanobacteria, particularly heterocyst-forming cyanobacteria, but also significant input of DNA from distantly related bacteria.

In these regards, the *N. punctiforme* and *Anabaena* PCC 7120 genomes are similar. The *Synechocystis* PCC 6803 genome shares certain features such as a large number of transposable elements and HIP1 sequences, but not others in that there are very few multilocus repeats and restriction/modification systems. All three, however, are strikingly different from the genomes of *P. marinus* MED4 and *Synechococcus* WH8102 which lack all of these elements. Perhaps these marine cyanobacteria are not often exposed to foreign DNA and thus have not modified their genomes to exploit new genetic opportunities.

The phenotype of *N. punctiforme* is complex, more so than most other cyanobacteria, and essentially all genes are present that might be required to support the multifaceted phenotype. The great breadth of tools available to *N. punctiforme* to sense and respond to environmental signals was not anticipated. Moreover, the identification of putative toxin synthetic genes, those for circadian rhythms, alternative environmental sources of phosphorous and sulfate, osmoregulation via glycine betaine uptake and Ccm indicate that nearly all characteristic that are collectively represented in cyanobacteria as a group are present in *N. punctiforme*. Apparently absent are genes involved in sulfide oxidation to support PS2-independent linear photosynthetic electron transport, analogous to that in anoxygenic green sulfur bacteria (present in a few cyanobacteria such as *Oscillatoria limnetica*; Oren 2000), rapid response to desiccation and rewetting (Pentecost and Whitton 2000), and perhaps Type III CCA.

Clearly, genes must be present that determine two of the most interesting phenotypic characteristics of *N. punctiforme*, cell differentiation and symbiotic interaction. The best understood examples of bacterial differentiation, such as sporulation by *Bacillus* and *Myxococcus* and swarmer and stalk cell formation by *Caulobacter*, provide a wealth of genes specific for these behaviors. However, ORFs in the *N. punctiforme* genome show no more similarity to these than to genes from nondifferentiating bacteria. Similarly, genes from Rhizobia known to be involved in their interaction with legumes have not been useful in identifying genes required for the interaction of *N. punctiforme* with plants. No doubt, the regulatory mechanisms governing complex bacterial behaviors evolved multiple times, drawing on the pool of signal transduction protein kinases and other regulatory proteins that are so abundant in *N. punctiforme*.

It will thus be necessary to rely on the traditionally genetic approaches of mutation and phenotypic characterization to define those genes necessary for the differentiation of akinetes, heterocysts and hormogonia, and for the ability to enter into symbiotic associations.

While many of these genes may have already been recognized as encoding regulatory proteins, many others may lie amongst the 2,164 hypothetical ORFs thus far unique to *N. punctiforme* and particularly amongst the 486 hypothetical genes shared by *N. punctiforme* and *Anabaena* PCC 7120. Within the latter group is a gene whose product positively regulates heterocyst differentiation (Wong and Meeks 2001), and two other genes whose products are important in establishing the pattern of heterocyst spacing (F.C. Wong and J.C. Meeks, unpublished).

The ability to manipulate the genome of *N. punctiforme* by sequence specific recombination-directed mutation and transposon mutagenesis (Cohen et al. 1998; Hagen and Meeks 1999) will prove invaluable in identifying the function of genes involved in the complex behaviors exhibited by this organism. The availability of the genomic sequence, moreover, offers the possibility of using global gene expression methodologies to identify genes transcribed under particular conditions. Genes identified in this way and fused to easily assayed reporters (Cohen and Meeks 1997; Wong and Meeks 2001) may permit the elucidation of one of the biggest prizes *N. punctiforme* has to offer: the mechanisms by which cyanobacteria and plants communicate with each other to yield a stable nitrogen fixing symbiosis.

Acknowledgements

The sequencing effort was supported by the DOE through contract with the Joint Genome Institute. Research in the laboratories of JM, JE, TT and MP is supported by grants from the DOE, NSF and USDA. We thank Elsie Campbell, Kari Hagen, John Ingraham and Francis Wong for sequence analyses and for critical review of the manuscript.

References

Adams DG and Duggan PS (1999) Heterocyst differentiation and akinete differentiation in cyanobacteria. New Phytol 144: 3-33

Angeloni SV and Potts M (1994) Analysis of the sequences within and flanking the cyanoglobin-encoding gene, *glbN*, of the cyanobacterium *Nostoc commune* UTEX 584. Gene 146: 133-134

Beale SI (1999) Enzymes of chlorophyll biosynthesis. Photosyn Res 60: 43-73

Bhaya D, Schwarz R and Grossman AR (2000) Molecular responses to environmental stress. In: Whitton BA and Potts M (eds) The Ecology of Cyanobacteria Their Diversity in Time and Space, pp 397-442. Kluwer Academic Publishers, Dordrecht, The Netherlands

Bickle T and Krüger DH (1993) Biology of DNA restriction. Microbiol Rev 57:434-450

Black K, Buikema W and Haselkorn R (1995) The *hglK* gene is required for localization of heterocyst-specific glycolipids in the cyanobacterium *Anabaena* sp. strain PCC 7120. J Bacteriol 177: 6440-6448

Black TA and Wolk CP (1994) Analysis of a Het$^-$ mutation in *Anabaena* sp. strain PCC 7120 implicates a secondary metabolite in the regulation of heterocyst spacing. J Bacteriol 176:2282-2292

Blattner FR, Plunkett III G, Bloch CA, Perna NT, Burland V and 12 others (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453-1479

Brusca JS, Hale MA, Carrasco CD and Golden JW (1989) Excision of an 11-kilobase-pair DNA element from within the *nifD* gene in *Anabaena variabilis* heterocysts. J Bacteriol 171: 4138-4145

Buikema WJ and Haselkorn R (1993) Molecular genetics of cyanobacterial development. Ann Rev Plant Physiol and Plant Mol Biol 44: 33-52

Cai Y and Wolk CP (1997) Nitrogen deprivation of *Anabaena* sp. strain PCC 7120 elicits rapid activation of a gene cluster that is essential for uptake and utilization of nitrate. J Bacteriol 179: 258–266

Campbell EL, Brahamsha B and Meeks JC (1998) Mutation of an alternative sigma factor in the cyanobacterium *Nostoc punctiforme* results in increased infection of its symbiotic plant partner, *Anthoceros punctatus*. J Bacteriol 180: 4938-4941

Campbell EL, Hagen KD, Cohen MF, Summers ML and Meeks JC (1996) The *devR* gene product is characteristic of receivers of two-component regulatory systems and is essential for heterocyst development in the filamentous cyanobacterium *Nostoc* sp. strain ATCC 29133 J Bacteriol 178: 2037-2043

Campbell EL and Meeks JC (1989) Characteristics of hormogonia formation by symbiotic *Nostoc* spp. in response to the presence of *Anthoceros punctatus* or its extracellular products App Environ Microbiol 55: 125-131

Campbell EL and Meeks JC (1992) Evidence for plant-mediated regulation of nitrogenase expression in the *Anthoceros-Nostoc* symbiotic association. J Gen Microbiol 138: 473-480

Carrasco CD, Buettner JA and Golden JW (1995) Programmed DNA rearrangement of a cyanobacterial *hupL* gene in heterocysts. Proc Natl Acad Sci USA 92: 791-795

Castenholz RW and Waterbury JB (1989) Oxygenic photosynthetic bacteria. Group I. cyanobacteria. In: Staley JT, Bryant MP, Pfenning N and Holt JG (eds) Bergey's Manual of Systematic Bacteriology, vol. 3, pp. 1710-1789. Williams and Wilkins, Baltimore.

Cohen MF and Meeks JC (1997) A hormogonium regulating locus *hrmUA*, of the cyanobacterium *Nostoc punctiforme* strain ATCC 29133 and its response to an extract of a symbiotic plant partner *Anthoceros punctatus*. Mol Plant-Microbe Inter 10: 280-289

Cohen MF, Meeks JC, Cai YA and Wolk CP (1998) Transposon mutagenesis of heterocyst-forming filamentous cyanobacteria. Meth Enzymol 297: 3-17

Curtis SE and Martin JA (1994) The transcription apparatus and the regulation of transcription initiation. In: Bryant DA (ed) The Molecular Biology of Cyanobacteria, pp 613-639. Kluwer Academic Publishers, Dordrecht, The Netherlands

Dean DR and Jacobson MR (1992) Biochemical genetics of nitrogenase. In: Stacey G, Burris RH and Evans HJ (eds) Biological Nitrogen Fixation, pp 763-834. Chapman and Hall, Inc. New York

Des Marais DJ (2000) When did photosynthesis emerge on earth? Science 289: 1703-1705 Douglas SE (1994) Chloroplast origins and evolution. In: Bryant DA (ed) The Molecular Biology of Cyanobacteria, pp 91-118. Kluwer Academic Publishers, Dordrecht, The Netherlands

Dow CS and Swoboda UK (2000) Cyanotoxins. In: Whitton BA and Potts M (eds) The Ecology of Cyanobacteria Their Diversity in Time and Space, pp 613-632. Kluwer Academic Publishers, Dordrecht, The Netherlands

Elhai J (2001) Determination of bias in the relative abundance of oligonucleotides in DNA sequences. J Comput Biol 8: 151-176

Enderlin CS and Meeks JC (1983) Pure culture and reconstitution of the *Anthoceros-Nostoc* symbiotic association. Planta 158: 157-165

Fiedler G, Arnold M and Maldener I (1998) Sequence and mutational analysis of the *devBCA* gene cluster encoding a putative ABC transporter in the cyanobacterium *Anabaena variabilis* ATCC 29413. Biochim Biophys Acta. 1375: 140-143

Field D and Wills C (1998) Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. Proc Natl Acad Sci USA 95: 1647-1652

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF and 35 others (1995) Whole genome shotgun sequencing and assembly of *Haemophilus influenza* Rd. Science 269: 496-512

Flores E and Herrero A (1994) Assimilatrory nitrogen metabolism and its regulation. In: Bryant DA (ed) The Molecular Biology of Cyanobacteria, pp 487-517. Kluwer Academic Publishers, Dordrecht, The Netherlands

Frías JE, Flores E and Herrero A (1994) Requirement of the regulatory protein NtcA for the expression of nitrogen assimilation and heterocyst development genes in the cyanobacterium *Anabaena* sp. PCC 7120. Mol. Microbiol 14: 823-832

Frías JE, Flores E and Herrero A (1997) Nitrate assimilation gene cluster from the heterocyst-forming cyanobacterium *Anabaena* sp. strain PCC 7120. J Bacteriol 179: 477-486

Fukumori F and Saint CP (1997) Nucleotide sequences and regulational analysis of genes involved in conversion of aniline to catechol in *Pseudomonas putida* UCC22 (pTDN1). J Bacteriol 179: 399-408

Garay-Arroyo A, Colmenero-Flores M, Garciarrubio A and Covarrubias AA (2000) Highly hydrophilic proteins in prokaryotes and eukaryotes are common during conditions of water deficit. J Biol Chem 275: 5668-5674

Golden JW, Carrasco CD, Mulligan ME, Schneider GJ and Haselkorn R (1988) Deletion of a 55-kilobase-pair DNA element from the chromosome during heterocyst differentiation of *Anabaena* sp. strain PCC 7120. J Bacteriol 170: 5034-5041

Golden JW, Robinson SJ and Haselkorn R (1985) Rearrangement of nitrogen fixation genes during heterocyst differentiation in the cyanobacterium *Anabaena*. Nature 314: 419-423

Golden SS (1994) Light-responsive gene expression and the biochemistry of the photosystem II reaction center. In: Bryant DA (ed) The Molecular Biology of Cyanobacteria, pp 693-714. Kluwer Academic Publishers, Dordrecht, The Netherlands

Golden SS, Johnson CH and Kondo T (1998) The cyanobacterial circadian system: a clock apart. Curr Opin Microbiol 1: 669-673

Hagen KD and Meeks JC (1999) Biochemical and genetic evidence for participation of *devR* in a phosphorelay signal transduction pathway essential for heterocyst maturation in *Nostoc punctiforme* ATCC 29133. J Bacteriol 181: 4430-4434

Hagen KD and Meeks JC (2001) The unique cyanobacterial protein OpcA is an allosteric effector of glucose-6-phosphate dehydrogenase in *Nostoc punctiforme* ATCC 29133. J Biol Chem 276: 11477-11486

Hebbar PB and Curtis SE (2000) Characterization of *devH*, a gene encoding a putative DNA binding protein required for heterocyst function in *Anabaena* sp. strain PCC 7120. J Bacteriol 182: 3572-8351

Hoch JA and Silhavy TJ (1995) Preface. In: Hoch JA and Silhavy TJ (eds) Two-Component Signal Transduction, p xv. American Society for Microbiology, Washington DC

Holland D and Wolk CP (1990) Identification and characterization of *hetA*, a gene that acts early in the process of morphological differentiation of heterocysts. J Bacteriol 172: 3131-3137

Huang TC, Lin RF, Chu MK and Chen HM (1999) Organization and expression of nitrogen-fixation genes in the aerobic nitrogen-fixing unicellular cyanobacterium *Synechococcus* sp. strain RF-1. Microbiology 145: 743-753

Hunsucker SW, Tissue BM, Potts M and Helm RF (2001) Screening protocol for the ultraviolet-photoprotective pigment scytonemin. Anal Biochem 288: 227-230

Ireton K, Rudner Z, Siranosian LKJ and Grossman AD (1993) Integration of multiple developmental signals in *Bacillus subtilis* through the Spo0A trascription factor. Genes Dev 7: 283-294

Ishiura M, Kutsuna S, Aoki S, Iwasaki H, Andersson CR, Tanabe A, Golden SS, Johnson CH and Kondo T (1998) Expression of a gene cluster *kaiABC* as a circadian feedback process in cyanobacteria. Science 281: 1519-1523

Johansson C and Bergman B (1994) Reconstitution of the symbiosis of *Gunnera manicata* Linden: cyanobacterial specificity. New Phytol 126: 643-652

Karlin S, Mrazek J, and Campbell AM (1997) Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179: 3899-3913

Khudyakov I and Wolk CP (1996) Evidence that the *hanA* gene coding for HU protein is essential for heterocyst differentiation in, and cyanophage A-4(L) sensitivity of, *Anabaena* sp. strain PCC 7120. J Bacteriol 178: 3572-3577

Khudyakov I and Wolk CP (1997) *hetC*, a gene coding for a protein similar to bacterial ABC protein exporters, is involved in early regulation of heterocyst differentiation in *Anabaena* sp. strain PCC 7120. J Bacteriol 179: 6971-6978

Kobayashi I, Nobusato A, Kobayashi-Takahashi N and Uchiyama I (1999) Shaping the genome-restriction-modification systems as mobile genetic elements. Curr Opin Genet Dev 9: 649-656

Kurnst F, Ogasawara N, Moszer I, Albertini AM, Alloni G and 156 others (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature 390: 249-256

Lavine E and Thiel T (1987) UV-inducible DNA repair in the cyanobacteria *Anabaena* spp. J Bacteriol 58: 511-522

Liang J, Scappino L and Haselkorn R (1993) The *patB* gene product, required for growth of the cyanobacterium *Anabaena* sp. strain PCC 7120 under nitrogen-limiting conditions, contains ferredoxin and helix-turn-helix domains. J Bacteriol 175: 1697-1704

Liang J, Scappino L and Haselkorn R (1992) The *patA* gene product, which contains a region similar to CheY of *Escherichia coli*, controls heterocyst pattern formation in the cyanobacterium *Anabaena* 7120. Proc Natl Acad Sci USA 89: 5655-5659

Lynn ME, Bantle JA and Ownby JD (1986) Estimation of gene expression in heterocysts of *Anabaena variabilis* by using DNA-RNA hybridization. J Bacteriol 167: 940-946

Masepohl B, Gorlitz K and Bohme H (1996) Long tandemly repeated repetitive (LTRR) sequences in the filamentous cyanobacterium *Anabaena* sp. PCC 7120. Biochem Biophys Acta 1307: 26-30

Matveyev AV, Young, KT, Meng, A and Elhai J (2001) DNA methyltransferases of the cyanobacterium *Anabaena* PCC 7120 Nucl Acid Res 29: 1491-1506

Mazel D, Houmard J, Castets AM and Tandeau de Marsac N (1990)

Highly repetitive DNA sequences in cyanobacterial genomes. J Bacteriol

172: 2755-2761

Meeks JC (1998) Symbiotic associations between nitrogen-fixing cyanobacteria and plants. BioScience 48: 266-276

Merrick MJ and Edwards RA (1995) Nitrogen control in bacteria. Microbiol Rev 59: 604-622

Metcalf WW and Wolfe RS (1998) Molecular genetic analysis of phosphite and hypophosphite oxidation by *Pseudomonas stutzeri* WM88. J Bacteriol 180: 5547-5558

Mollenhauer D, Mollenahuer R and Kluge M (1996) Studies on initiation and development of the partner association in *Geosiphon pyriforme* (Kutz.) v. Wettstein, a unique encocytobiotic system of a fungus (Glomales) and the cyanobacterium *Nostoc punctiforme* (Kutz.). Hariot Protoplasma 193: 3-9

Morand, LZ, Cheng RH, Krogmann DW and Ho KK (1994) Soluble electron transfer catalysts of cyanobacteria. In: Bryant DA (ed) The Molecular Biology of Cyanobacteria, pp 381-407. Kluwer Academic Publishers, Dordrecht, The Netherlands

Naylor GW, Addlesee HA, Gibson LCD and Hunter CW (1999). The photosynthesis gene cluster of *Rhodobacter sphaeroides*. Photosyn Res 62: 121-139

Newman J, Karakaya H, Scanlan DJ and Mann NH (1995) A comparison of gene organization in the *zwf* region of the genomes of the cyanobacteria *Synechococcus* sp. PCC 7942 and *Anabaena* sp. PCC 7120. FEMS Lett 133: 187-193

Okuhara H, Matsumura T, Fujita Y and Hase T (1999) Cloning and inactivation of genes encoding ferredoxin- and NADH-dependent glutamate synthases in the cyanobacterium *Plectonema boryanum*.

Imbalances in nitrogen and carbon assimilations caused by deficiency of the ferredoxin-dependent enzyme. Plant Physiol 120: 33-42

Oren A (2000) Salts and Brines. In: Whitton BA and Potts M (eds) The Ecology of Cyanobacteria Their Diversity in Time and Space, pp 281-306. Kluwer Academic Publishers, Dordrecht, The Netherlands

Pentecost A and Whitton BA (2000) Limnestones. In: Whitton BA and Potts M (eds) The Ecology of Cyanobacteria Their Diversity in Time and Space, pp 257-279. Kluwer Academic Publishers, Dordrecht, The Netherlands

Potts M, Angelon SV, Ebel RE and Bassam D (1992) Myoglobin in a cyanobacterium. Science 256: 1690–1692

Ramaswamy KS, Carrasco CD, Fatma T and Golden JW (1997) Cell-type specificity of the *Anabaena fdxN*-element rearrangement requires *xisH* and *xisI*. Mol Microbiol 23: 1241-1249

Rippka R and Herdman M (1992) Pasteur Culture Collection of Cyanobacterial Strain in Axenic Culture. Institut Pasteur, Paris, France

Robinson NJ, Rutherford JC, Pocock MR and Cavet JS (2000) Metal metabolism and toxicity: repetitive DNA. In: Whitton BA and Potts M (eds) The Ecology of Cyanobacteria Their Diversity in Time and Space, pp 443-463. Kluwer Academic Publishers, Dordrecht, The Netherlands

Sakamoto T, Inoue-Sakamoto K and Bryant DA (1999) A novel nitrate/nitrite permease in the marine cyanobacterium *Synechococcus* sp. strain PCC 7002. J Bacteriol 181: 7363-7372.

Sato N (1994) A cold-regulated cyanobacterial gene cluster encodes RNA-binding protein and ribosomal protein S21. Plant Mol Biol 24: 819-823

Schmetterer G (1994) Cyanobacterial respiration. In: Bryant DA ed) The Molecular Biology of Cyanobacteria, pp 409-435. Kluwer Academic Publishers, Dordrecht, The Netherlands

Schmitz O, Katayama M, Williams SB, Kondo T and Golden SS (2000) CikA, a bacteriophytochrome that resets the cyanobacterial circadian clock. Science 289: 765-768

Schopf JW (2000) The fossil record: tracing the roots of the cyanobacterial lineage. In: Whitton BA and Potts M (eds) The Ecology of Cyanobacteria Their Diversity in Time and Space, pp13-35. Kluwer Academic Publishers, Dordrecht, The Netherlands

Sitachitta N, Marquez BL, Williamson RT, Rossi J, Roberts MA, Gerwick WH, Nguyen V-A and Willis CL (2000) Biosynthetic pathway and origin of the chlorinated methyl group in barbamide and dechlorobarbamide, metabolites from the marine cyanobacterium *Lyngbya majuscula*. Tetrahedron 56: 9103-9113

Smith AJ (1982) Modes of cyanobacterial carbon metabolism. In: Carr NG and Whitton BA (eds) The Biology of Cyanobacteria, pp 47-85. Blackwell Scientific Publications, Oxford, UK

Stover CK, Pham XQ, Ersin AL, Mizoguchi, SD and 28 others (2000) Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. Nature 406: 959-964

Summers ML, Wallis JG, Campbell EL and Meeks JC (1995) Genetic evidence of a major role for glucose-6-phosphate dehydrogenase in nitrogen fixation and dark growth of the cyanobacterium *Nostoc* sp. strain ATCC 29133. J Bacteriol 177: 6184-6194

Tamagnini P, Costa JL, Almeida L, Oliveira MJ, Salema R and Lindblad P (2000) Diversity of cyanobacterial hydrogenases, a molecular approach. Curr Microbiol 40: 356-361

Tandeau de Marsac N (1977) Occurrence and nature of chromatic adaptation in cyanobacteria. J. Bacteriol 130: 82-91

Tandeau de Marsac N and Houmard J (1993) Adaptation of cyanobacteria to environmental stimuli: new steps towards molecular mechanisms FEMS Microbiol Rev 104: 119-190

Thiel T (1996) Isolation and characterization of the *vnfEN* genes of the cyanobacterium *Anabaena variabilis*. J Bacteriol 178: 4493-4499

Thiel T, Lyons EM and Erker JC (1997) Characterization of genes for a second Mo-dependent nitrogenase in the cyanobacterium *Anabaena variabilis*. J Bacteriol 179: 5222-5225

Thiel T, Lyons EM and Thielemeier J (1998) Organization and regulation of two clusters of *nif* genes in the cyanobacterium *Anabaena variabilis*. In: Peschek GA, Loeffelhardt W and Schmetterer G (eds) Phototrophic Prokaryotes. pp 517-521. Plenum Press, N.Y.

Thiel T and Pratte B (2001) Effect on heterocyst differentiation of nitrogen fixation in vegetative cells of the cyanobacterium *Anabaena variabilis* ATCC 29413. J Bacteriol 183: 280-286

Wang Q, Li H and Post AF (2000) Nitrate assimilation genes of the marine diazotrophic, filamentous cyanobacterium *Trichodesmium* sp. strain WH9601. J Bacteriol 182: 1764-1767

Wolk CP (2000) Heterocyst formation in *Anabaena*. In: Braun Y and Shimkets LJ (eds) Prokaryotic Development, pp 83-103. American Society of Microbiology, Washington DC

Wolk CP, Ernst E and Elhai J (1994) Heterocyst metabolism and development. In: Bryant DA (ed) The Molecular Biology of Cyanobacteria, pp 769-823. Kluwer Academic Publishers, Dordrecht, The Netherlands

Wong FCY and Meeks JC (2001) The *hetF* gene product is essential to heterocyst differentiation and affects HetR function in the cyanobacterium *Nostoc punctiforme*. J Bacteriol 183: 2654-2661

Wright D, Prickett T, Helem RF and Potts, M (2001) Form species *Nostoc commune* (Cyanobacteria). Int J Sys Evol Microbiol 51: 1839-1852

Yeh K-O, Wu S-H, Murphy JT and Lagarias JC (1997) A cyanobacterial phytochrome two-component light sensory system. Science 277: 1505-1508

Yoon HS and Golden JW (1998) Heterocyst pattern formation controlled by a diffusible peptide. Science 282: 935-938

Zhou R, Wei X, Jiang N, Li H, Dong Y, Hsi KL and Zhao J (1998) Evidence that HetR protein is an unusual serine-type protease. Proc Natl Acad Sci USA 95: 4959-4963

Zhu J, Kong R and Wolk CP (1998) Regulation of *hepA* of *Anabaena* sp. strain PCC 7120 by elements 5' from the gene and by *hepK*. J Bacteriol 180: 4233-4242