Structural analysis of a non-ribosomal halogenated cyclic peptide and its putative operon from *Microcystis*: implications for evolution of cyanopeptolins

Ave Tooming-Klunderud,1 Thomas Rohrlack,2 Kamran Shalchian-Tabrizi,3 Tom Kristensen1 and Kjetill S. Jakobsen3

Correspondence
Kjetill S. Jakobsen
kjetill.jakobsen@bio.uio.no

1Department of Molecular Biosciences, University of Oslo, 0316 Oslo, Norway
2NIVA, Norwegian Institute for Water Research, 0411 Oslo, Norway
3Department of Biology, Centre for Ecological and Evolutionary Synthesis (CEES), University of Oslo, 0316 Oslo, Norway

The structure of the major peptide produced by *Microcystis* cf. *wesenbergii* NIVA-CYA 172/5, the halogenated heptapeptide cyanopeptolin-984, was determined using LC/MS/MS. A gene cluster encoding a peptide synthetase putatively producing a cyanopeptolin was cloned from the same strain and sequenced. The cluster consists of four genes encoding peptide synthetases and one gene encoding a halogenase. Two additional ORFs transcribed in the opposite direction were found in the 5’ flanking sequence; one of these encodes an ABC transporter. The overall organization of the cyanopeptolin synthetase operon (*mcn*) resembles a previously analysed anabaenopeptilide synthetase operon (*apd*) from *Anabaena* strain 90. Phylogenetic analyses of the individual domains from Mmcn, Apd and other cyanobacterial peptide synthetases showed clustering of the adenylation domains according to function irrespective of operon origin – indicating strong functional constraints across peptide synthetases. In contrast, the condensation and thiolation domains to a large extent grouped according to operon affiliation or position in the respective operons. Phylogenetic analyses of condensation domains indicated that N-terminal domains and domains that condense L-amino acids and D-amino acids, respectively, form three separate groups. Although recombination events are likely to be involved in the evolution of *mcn*, no clear evidence of genetic recombination between the two cyanopeptolin gene clusters was found. Within the genus *Microcystis*, microcystin and cyanopeptolin synthetases have an evolutionary history of genomic coexistence. However, the data indicated that the two classes of peptide synthetase gene clusters have evolved independently.

INTRODUCTION

Cyanobacteria produce a large number of bioactive cyclic peptides, both toxic and non-toxic (Namikoshi & Rinehart, 1996). The cyclic structure and presence of unusual and modified amino acids indicate that many of these compounds are synthesized by non-ribosomal peptide synthetases (NRPSs) using the thiotemplate mechanism (Marahiel et al. 1997). NRPSs are large multienzyme complexes organized in iterated modules, one module for each amino acid incorporated into the peptide. The minimal module required for the addition of an amino acid to the growing peptide consists of a condensation (C), an adenylation (A) and a thiolation (T) domain. The A-domain is involved in the selection and activation of the amino acid substrate, which then becomes covalently attached to the enzyme via a thioester bond to the phosphopantetheine group linked to the T-domain. The C-domain catalyses the formation of a peptide bond between the aminoacyl or peptidyl moiety and the free amino group of the downstream aminoacyl moiety. The more sporadic presence of epimerization (E), N-methyltransferase (NMT), cyclization and oxidation domains contributes to the structural diversity of non-ribosomal peptides (see Sieber & Marahiel, 2005).

Despite the large number of cyclic peptides found in cyanobacteria, relatively few classes of NRPS operons have...
been characterized to date. The most intensively studied so far are the microcystin class of synthetase gene clusters (mcy, nda), which have been sequenced from two strains of Microcystis (Nishizawa et al., 2000; Tillett et al., 2000), Planktothrix agardhii (Christiansen et al., 2003) Anabaena strain 90 (Rouhiainen et al., 2004) and Nodularia spumigena (Moffitt & Neilan, 2004). The synthetases encoded by these operons contain both NRPS and polyketide synthetase (PKS) modules. Other examples of hybrid biosynthesis between NRPS and PKS are nostopeptolide and nostocyclopeptide synthetases from Nostoc spp. (Becker et al., 2004; Hoffmann et al., 2003). Many cyanobacterial groups, particularly Microcystis strains, produce other cyclic and linear peptides, including cyanopeptolins, microvirdins and aeruginosins (Birk et al., 1989; Ishitsuka et al., 1990; Murakami et al., 1997). So far only a single ‘pure’ (without PKS modules) NRPS gene cluster, encoding a seven-residue depsipeptide (anabaenopeptide) belonging to the cyanopeptolins family, has been characterized from Anabaena (Rouhiainen et al., 2000).

The modular structure of NRPS gene clusters, their diversity and their patchy distribution among cyanobacterial species may hint at recombination and lateral gene transfer (LGT) as key processes in the evolution of these gene clusters. Phylogenetic studies of prokaryotic A-domains have shown that they essentially cluster according to function (i.e. according to the amino acid activated) (Challis et al., 2000). For Bacillus subtilis it has been proposed that LGT of large NRPS gene clusters between strains is responsible for the very similar gene clusters present (Tsuge et al., 2001). Investigating the sequence variation in mcy genes of various Microcystis strains, Mikalsen et al. (2003) suggested that frequent recombination between imperfect repeats (i.e. domains) as well as gene loss/deletions and LGT events have contributed to the variation seen for the mcyABC operon within different strains. Recombination events, possibly involving intragenomic recombinations and/or LGT events, were recently suggested for mcy genes in Planktothrix (Christiansen et al., 2003) and for the mcy-like nda genes of N. spumigena (Moffitt & Neilan, 2004). Tanabe et al. (2004) inferred a reticulate phylogeny of mcyA and identified several potential recombination tracts of mcyA, but none in the PKS part (mcyD, G and J) of the gene cluster. Notably, transposases are found to be associated with all mcy gene clusters characterized as well as with the nda gene cluster (Christiansen et al., 2003, 2006; Moffitt & Neilan, 2004; Rouhiainen et al., 2004; Tillett et al., 2000). The high number of isoforms and the recent discoveries of frequent recombinations (Kurmayer et al., 2004; Mikalsen et al., 2003; Tanabe et al., 2004) may indicate a rapid evolution of mcy genes. In contrast, the findings of Rantalä et al. (2004), suggesting that microcystin synthetase genes were present in the last common ancestor of a deep radiation of cyanobacteria, are consistent with an evolutionary process that acts in preserving the basic structure of the peptides, enzymes and operons.

The genus Microcystis represent an evolutionary lineage synthesizing both microcystin and other peptides, and co-production of microcystin and cyanopeptolin has been shown in several strains (see Birk et al., 1989; Ishida et al., 1995, 1998; Martin et al., 1990; Murakami et al., 1997; Yasuno et al., 1998). An interesting question then is what happens over evolutionary time when two or more NRPS gene clusters coexist in the same genome. Will they prevail as ‘insulated islands’, or will various recombination mechanisms gradually turn them into mosaics? One way of addressing this issue is to investigate a non-microcystin operon in a microcystin-producing lineage, preferably an operon from a class with some other genetically investigated members. Using MS screening, we have found that the strain Microcystis cf. wesenbergii NIVA-CYA 172/5 (Microcystis N-C 172/5) produces a cyanopeptolin. Cyanopeptolins are a class of protease-inhibiting peptides characterized by the presence of the amino acid 3-amino-6-hydroxy-2-piperidone (Ahp) and the cyclization of the peptide by an ester bond between the β-hydroxy group of threonine and the carboxy group of the terminal amino acid (Martin et al., 1993). Next to microcystins, cyanopeptolins appear to be the most common peptide produced by Microcystis species and were detected in 60 % of single Microcystis colonies in a recent survey (Welker et al., 2006). Gene clusters responsible for the production of these two peptide families therefore must coexist in many Microcystis species, as they do for example in Anabaena strain 90 (Rouhiainen et al., 2004). Further information about cyanobacterial peptides and NRPSs, including cyanopeptolins, can be found in a recent review (Welker & von Döhren, 2006). We have determined the structure of the N-C 172/5 peptide. In addition we have sequenced a complete cyanopeptolin operon and its flanking sequences and addressed the evolution of the cyanopeptolin gene clusters from Microcystis and Anabaena in a wider context to gain insight into the evolutionary mechanisms that shape the combinational nature of peptide synthetase operons.

### METHODS

**Cyanobacterial strain and growth conditions.** Microcystis cf. wesenbergii NIVA-CYA 172/5 (N-C 172/5), isolated from Lake Arresø, Denmark, was cultured at the Norwegian Institute for Water Research (NIVA) as previously described by Skulberg & Skulberg (1990).

**Chemical analyses.** The chemical analysis of Microcystis N-C 172/5 included an initial search for a compound with the anticipated molecular mass and a subsequent structural elucidation. In both cases a methanol extract of lyophilized N-C 172/5 cells was used as basis. The initial screening was done by LC/MS using a Waters LC system (600 S controller, Waters 626 pump, Waters Symmetry column 2.1 × 150 mm, particle size 5 μm) set to run a gradient starting with 100 % solvent A (84.9 % water, 15 % acetonitrile, 0.1 % acetic acid) and ending with 100 % solvent B (80 % acetonitrile, 19.9 % water, 0.1 % acetic acid) after 30 min. The flow rate was 0.5 ml min⁻¹. The LC system was connected through a 1:10 flow splitter to a Micromass Platform 2 MS detector equipped with an electrospray probe. The detector was run in the positive ion mode at a cone voltage of 50 V. A total ion scan from 500 to 2000 Da was performed during the entire length of the LC gradient. Once a candidate compound had been identified, its structure was analysed by
MS fragmentation using a ThermoFinnigan LC/MS/MS system (Surveyor LC solvent delivery system at a flow rate of 0.2 ml min⁻¹, TS quantum MS/MS detector). The LC column and gradient were identical to those described above. The MS/MS detector was run in positive electrospray ionization mode. The candidate compound was set as parent molecule. All settings of the source, collision cell as well as quadrupoles 1 and 3 were optimized automatically at collision energies of 30, 40, 50 and 60 eV. Fragments with molecular masses of 30 to 1000 Da were recorded during the entire length of the LC gradient. The fragmentation spectra were screened for ion-molecule ions of amino acids and other typical small indicative fragments to be expected when fragmenting peptides. Larger fragments were identified by comparison with a fragment library of purified cyanobacterial oligopeptide maintained at NIVA. The identification was further assisted by comparison with theoretical fragment spectra calculated based on the structural predictions of the genetic analyses. Finally, all information on fragments was used to gradually develop a model of the compound’s structure. This method of structural elucidation has been previously used extensively and successfully for cyanobacterial oligopeptides (e.g. Fastner et al., 2001; Welker et al., 2006).

The amounts of Ile and Leu in the peptide were determined by means of an amino acid analyser (model 421, Applied Biosystems) after acid hydrolysis.

**DNA isolation and manipulations.** A Wizard Genomic DNA Purification Kit (Promega) was used for isolation of genomic DNA. For preparation and transformation of competent Escherichia coli cells, digestion and ligation of DNA, amplification and purification of plasmids, standard procedures were used (Sambrook & Russell, 2001).

**Identification of the putative cyanopeptolin biosynthetic gene cluster and its flanking regions.** Genes encoding putative peptide synthetases were identified using consensus primers derived from highly conserved regions in NRPS modules, followed by primer walking. To screen the Microcystis strain for peptide synthetase A-domains, we performed PCR with the primer pair MTF2 and MTR (Nelän et al., 1999) and BD Advantage 2 polymerase (BD Biosciences). The resulting amplicons were cloned in the vector pGEM-T-Easy (Promega), and 23 random clones were sequenced. Three different peptide synthetase gene fragments were obtained, and the rest of the gene cluster was amplified using specific PCR and hemidegenerate PCR. In the hemidegenerate PCR approach, specific PCR primers were used together with degenerate A-domain PCR primers (MTF2 and MTF; Nelän et al., 1999). For amplification of flanking regions, we performed PCR using a single, specific primer and a low annealing temperature (Malo et al., 1994). This approach successfully generated a series of overlapping fragments. To avoid possible artefacts introduced during PCR, several pGEM-T-Easy clones were sequenced for each PCR. DNA sequencing was performed on both strands using a DYEnamic ET Dye Terminator Cycle Sequencing Kit (GE Healthcare) and an automated DNA sequencer (MegaBACE 1000, GE Healthcare). The nucleotide sequence of the cyanopeptolin synthetase gene cluster described in this paper has been submitted to GenBank under accession number DQ075244.

**RNA extraction and reverse transcription-PCR (RT-PCR).** Total RNA was extracted from 50 ml culture as described previously (Kaebernick et al., 2000). After extraction with Trizol (Invitrogen), RNA was column purified using the SV total-RNA Isolation System (Promega). DNA was removed by treatment with DNase I (Fermentas). Reverse transcription for production of cDNA for RT-PCR was carried out using the SuperScript III First-Strand Synthesis System for RT-PCR (Invitrogen). The presence of cyanopeptolin synthetase transcripts was verified by RT-PCR. Four different primer pairs from different parts of the gene cluster were used: mcnA (mcnA-F, 5'-ACTCCATAATCTATGACTCGTAG-3'; mcnA-R, 5'-TACGGAACTCTCGATTTTACTTG-3'), mcnC (mcnC-F, 5'-GATTTTACGAGAATGATACG-3'; mcnC-R, 5'-ACGGGTTATAACTGCTAAAAACTCG-3'), mcnD, encoding a homologase (mcnD-F, 5'-TTGTCCAGAGTGATACGCACAG-3'; mcnD-R, 5'-ATAAACTCTGCTTCACAATTCC-3'), and mcnF, encoding an ABC transporter (mcnF-F, 5'-TTTTATTTGTGATATCTTGAC-3'; mcnF-R, 5'-CAGGAAATCTTTTACATC-3'). PCR without reverse transcription was used to check for DNA contamination in the total RNA used in RT-PCR.

**Phylogenetic analysis of A-, T- and C-domains.** Amino acid sequences encoded by A-, T- and C-domains from the mcn gene cluster were aligned with publicly available sequences using CLUSTALW (Thompson et al., 1997) and manual editing. After discarding ambiguously aligned characters, the A-domain data consisted of 352 characters and 40 sequences, the C-domain data of 300 characters and 39 sequences, and the T-domain data of 95 characters and 41 sequences. The sequences from the A-, T- and C-domains were used to infer the phylogeny in a Bayesian framework applying the program MrBayes v2.01, v3.0 and v3.1 (Ronquist & Huelsenbeck, 2003) on the freely available Bioportal computer resources website (http://www.bioportal.uio.no). Variable substitution rates across sites were accounted for by allowing each site to evolve with a rate defined as a random variable drawn from a gamma distributed rate probability (G), which was approximated using four discrete rate categories. The gamma distribution was used together with a category of invariant sites (I), together: G+I). In addition, the WAG stationary amino acid substitution model was used for A-, C- and T-domain datasets. Priors for all other model parameters were set to default values. Metropolis coupling was used with three heated (temperature parameter 0.2) and one cold chain. The Markov chain Monte Carlo (MCMC) chains started from random trees and lasted for 2 000 000–4 000 000 generations. Sampling of trees was done every 100 generations. Burn-in of trees was set to 5000 trees based on assessment of the likelihood plots. The consensus of the remaining trees was used to calculate the posterior probabilities of the clades. Two separate runs were performed for each dataset, and the posterior probability and likelihood of the trees were compared to confirm the stationarity of the MCMC chains. DNA sequences from all datasets were used to construct neighbour joining (NJ) distance trees using Kimura two-parameter and uniform rates among the sites. The statistical support for the internal branch topologies was analysed with bootstrapping using 1000 pseudoreplicates.

**RESULTS**

**Identification of a cyanopeptolin in Microcystis by MALDI-TOF and LC-MS**

In a previous study (Mikalsen et al., 2003) we investigated 16 Microcystis strains for the presence of microcystin isoforms using MALDI-TOF. This study also revealed that all strains contained additional peptides. Most of these peptides could not be identified by the MALDI-TOF approach. Some strains, including N-C 31, 161/1, 228/1, 264 and 172/5, contained putative cyanopeptolins (data not shown). Of these strains, N-C 172/5 did not produce microcystin. Using LC/MS the initial screening for low-molecular-mass compounds displayed one major peak with strong signals at 966, 984, 1006 and 1022 Da [M+H]^+ (see Supplementary Fig. S1, available with the online version of
this paper). Such a cluster is typical for cyanopeptolins; frequent loss of $\text{H}_2\text{O}$ (966 Da $\text{[M+H-H}_2\text{O}]^+$) and formation of $\text{Na}^+$ and $\text{K}^+$ adducts (1006 Da $\text{[M+Na}]^+$ and 1022 Da $\text{[M+K}]^+$) will occur under the MS conditions applied in this study. The signal of the original molecule at 984 Da $\text{[M+H]}^+$ was also within the usual mass range of cyanopeptolins. The unique isotope pattern suggested the occurrence of one Cl in the molecule.

**Determination of the cyanopeptolin structure by chemical analyses**

The signal at 984 Da identified by peptide screening was selected for fragmentation experiments, the results of which are shown in Table 1. To obtain a model of the structure, the fragments were gradually identified and combined, beginning from the lower end of the mass scale. The occurrence of the typical Ahp fragment at 114 Da supported the identification of the compound as cyanopeptolin. Typical amino acid fragments and immonium ions suggested the occurrence of Gln, Tyr, Thr and Ile or Leu in the molecule. Fragments with a mass of 127 and 198 Da implied a chlorination of Tyr, which is not unusual in cyanopeptolins (Welker & von Döhren 2006). However, the actual position of Cl within the Tyr ring could not be elucidated. Besides being chlorinated, Tyr was also methylated at its amino group (183.1 Da fragment). A number of larger fragments indicated that the cyanopeptolin ring had the amino acid sequence Thr-Gln-ClMeTyr-Phe-Ahp-Ile(or Leu). Other fragments (e.g. 254.1, 243.1, 195.1, 72.0 and 70.0 Da) implied that the side chain consisted of acetic acid bound to a Gln residue. Finally, the amino acid analysis confirmed that the cyanopeptolin contains Ile or allo-Ile and not Leu (data not shown). Taken together, the chemical analyses strongly suggested that *Microcystis* N-C 172/5 produces a cyanopeptolin with a molecular mass of 984 Da $\text{[M+H]}^+$ and a structure as shown in Fig. 1(a). We name the peptide cyanopeptolin-984 (cypep-984).

**Identification of a cyanopeptolin synthetase gene cluster**

A contiguous 31 880 bp region was fully sequenced on both strands, revealing a cluster of five putative ORFs transcribed in the same direction and showing a high degree of similarity to a number of bacterial NRPSs or associated proteins. The gene cluster characterized in the present study shows most overall similarity to the anabaenopeptilide synthetase gene cluster (apd) from *Anabaena* strain 90, with 52–93 % similarity between modules on the amino acid level (Fig. 2).

**Table 1.** LC/MS/MS fragmentation pattern of cyanopeptolin-984 produced by *Microcystis* N-C 172/5

<table>
<thead>
<tr>
<th>Collision energy</th>
<th>Fragment mass (Da)</th>
<th>Fragment(s)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 eV</td>
<td>984.4</td>
<td>$\text{M+H}^+$</td>
</tr>
<tr>
<td></td>
<td>966.4</td>
<td>$\text{M+H}^+ - \text{H}_2\text{O}$</td>
</tr>
<tr>
<td></td>
<td>797.3</td>
<td>$\text{M+H}^+ - \text{side chain} {\text{Gln-ClMeTyr-Phe-Ahp-Ile-Thr-NH}_3 + \text{H}^+}$</td>
</tr>
<tr>
<td></td>
<td>777.3</td>
<td>$\text{M+H}^+ - \text{side chain-H}_2\text{O}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>${\text{Gln-ClMeTyr-Phe-Ahp-Ile-Thr-NH}_3 - \text{H}_2\text{O} + \text{H}^+}$</td>
</tr>
<tr>
<td></td>
<td>454.1</td>
<td>Ahp[ring]-Phe-ClMeTyr-NH$_2$</td>
</tr>
<tr>
<td></td>
<td>453.1</td>
<td>Phe-ClMeTyr-Gln + $\text{H}^+ - 2\text{NH}_3$</td>
</tr>
<tr>
<td>40 eV</td>
<td>254.1</td>
<td>Ac-Gln-NH-C$_3$(+)H$_4 = C = O {\text{Ac-Gln-Thr-H}_2\text{O} + \text{H}^+}$</td>
</tr>
<tr>
<td></td>
<td>243.1</td>
<td>Ac-Gln-NH-C$_2$(+)H$_4$-CO-NH$_2$</td>
</tr>
<tr>
<td></td>
<td>198.0</td>
<td>NH$_2$-ClMeTyr-CH$_3$NH</td>
</tr>
<tr>
<td></td>
<td>159.1</td>
<td>OH-C(+)H-Ile-NH$_2$, OH-C(+)H-Ahp</td>
</tr>
<tr>
<td>50 eV</td>
<td></td>
<td>No additional fragments</td>
</tr>
<tr>
<td>60 eV</td>
<td>342.1</td>
<td>ClMeTyr-Phe+$\text{H}^+ - \text{NH}_2$</td>
</tr>
<tr>
<td></td>
<td>326.2</td>
<td>Gln-Thr-Ile-NH$_2 + \text{H}^+ - 2\text{NH}_3$</td>
</tr>
<tr>
<td></td>
<td>214.2</td>
<td>CH$_3$-C(+)H-CHNH$_2$-CO-NH-Ile-NH$_2$</td>
</tr>
<tr>
<td></td>
<td>195.1</td>
<td>NH$_2$CO-CH$_2$-CH$_2$=CH-CO-NH-C$_3$(+)H$_4 = C = O {\text{from Gln-Thr}}$</td>
</tr>
<tr>
<td></td>
<td>183.1</td>
<td>Ahp[ring]-C(+)H-CO-NH-CH$_3$</td>
</tr>
<tr>
<td></td>
<td>127.0</td>
<td>CTyr ring</td>
</tr>
<tr>
<td>114.1</td>
<td></td>
<td>Ahp ring</td>
</tr>
<tr>
<td>101.1</td>
<td></td>
<td>N(+)H=CH-CH$_2$-CH$_2$-CONH$_2$ ${\text{Gln-CO} + \text{H}^+}$</td>
</tr>
<tr>
<td>97.1</td>
<td></td>
<td>CH$_3$-CH$_2$-CHCH$_2$-C(+) = C = O ${\text{from Ile}}$</td>
</tr>
<tr>
<td>84.1</td>
<td></td>
<td>CH$_3$-C(+)H-CHNH$_2$ = C = O ${\text{from Thr}}$</td>
</tr>
<tr>
<td>72.0</td>
<td></td>
<td>Ac-NH-C(+)H$_2$</td>
</tr>
<tr>
<td>70.0</td>
<td></td>
<td>Ac-N(+)CH</td>
</tr>
</tbody>
</table>

*Text in square brackets gives more detailed information on the identity of certain fragments.*
No NRPS modules apart from those present in this gene cluster were identified in the Microcystis strain using PCR with consensus primers.

The first ORF, \textit{mcnA}, is preceded by a putative promoter region. A putative \textit{E. coli} \textit{a}^{70}-like −10 element (TATATG, consensus TATAAT) was identified 133 bp upstream of the ATG start codon. The corresponding −35 element is less conserved (TTGCGA, consensus TTGACA). The second ORF (\textit{mcnB}) has two putative start codons: one ATG codon is located 3 bp and the other 15 bp downstream of the preceding TAA stop codon of the \textit{mcnA}. The third ORF (\textit{mcnC}) overlaps \textit{mcnB} by 4 bp and starts with an ATG start codon, while the ATG start codon of the fourth ORF (\textit{mcnD}) is located 4 bp upstream of the TGA stop codon of \textit{mcnC}. The fifth ORF (\textit{mcnE}) has two possible start codons: the first is GTG and is located 4 bp upstream of the TGA stop codon of \textit{mcnD}; the second is an ATG codon 93 bp downstream of the stop codon of \textit{mcnD}.

Analysis of the amino acid sequences of the \textit{mcn} gene products revealed seven modules typical of peptide synthetases: one in McnA, one in McnB, four in McnC and one in McnE. All modules contain a C-, an A- and a T-domain (Fig. 2a). The presence of a C-domain at the N-terminal end of the peptide synthetase is unusual, since the first module in the NRPS usually starts with an A-domain. However, a similar arrangement has been observed in nostopeptolide A synthetase (Nos) of \textit{Nostoc} sp. GSV224 (Hoffmann \textit{et al}., 2003) and lichenycin synthetase of \textit{Bacillus licheniformis} (Konz \textit{et al}., 1999). In these two systems, the N-terminal C-domain appears to be involved in N-acylation of the N-terminal amino acid.

The fourth module in McnC contains an insert of 427 amino acids between conserved regions A8 and A9 in the A-domain. This domain has extensive similarity to \textit{N}-methyltransferase (NMT) domains from different cyanobacterial peptide synthetases, especially to the NMT-domain in module 4 of ApdB in anabaenopeptilide synthetase, with 70 \% identity over 427 amino acids. The last module contains a thioesterase (TE) domain located at the C-terminal end of the McnE, and its function is most likely to offload the linear peptide intermediate from the NRPS and catalyse its cyclization. A BLAST search with this TE-domain found similarities to other NRPS TE-domains, especially to ApdD, with 76 \% identity over 282 amino acids, and to NosD, with 49 \% identity over 285 amino acids.

The fourth ORF, \textit{mcnD}, encodes a 625 amino acid polypeptide with no similarity to peptide synthetases. We found two highly conserved motifs characteristic of FADH$_2$-dependent halogenases (see Supplementary Fig. S2): a nucleotide binding motif GxGxxG (Scrutton \textit{et al}., 1990) and a motif WxWxIP, which might be involved in halide binding (van Pee & Unversucht, 2003). The amino acid sequence of McnD shows striking similarity to the halogenase ApdC, with 86 \% identity over 625 amino acids, while other halogenases associated with secondary metabolite biosynthetic gene clusters (Chiu \textit{et al}., 2001; Hammer \textit{et al}., 1997; Nowak-Thompson \textit{et al}., 1999; Pelzer \textit{et al}., 1999) show only 12–15 \% identity.

To confirm that the peptide synthetase cluster is active, the presence of cyanopeptolin synthetase transcripts was demonstrated by RT-PCR. Four different primer pairs from different parts of the gene cluster were used, and all gave positive results in the RT-PCR analyses; negative results when reverse transcription was omitted indicated no significant DNA contamination (data not shown).

**Peptide structure information gained from operon structure and the derived amino acid sequence**

Having obtained the entire operon at the DNA and protein sequence level and assuming that the encoded NRPS belongs to the linear class of NRPSs (Finking & Marahiel, 2004), we could deduce information about the resulting cyclic peptide...
structure using general knowledge about the modular structure of NRPSs, similarities between the A-domains in the operon and A-domains from the sequence databases, and the ‘binding pocket’ theory of Stachelhaus et al. (1999). The presence of seven modules in the Mcn sequence indicates that the peptide consists of seven amino acid residues. The seven A-domains of the mcn gene product were compared with database sequences using phylogenetic analyses (see details below), and also aligned with the Phe-activating A-domain of GrsA (Conti et al., 1997) to identify the ‘binding-pocket’ residues. Based on the comparisons and the binding pocket signatures, the amino acids likely to be activated by A-domains of different modules are shown in Table 2. According to this prediction, the A-domain in

Table 2. Deduced functions of encoded proteins in the mcn gene cluster of Microcystis N-C 172/5

<table>
<thead>
<tr>
<th>Protein</th>
<th>Length (aa)</th>
<th>Proposed domains</th>
<th>Putative substrate-binding pocket</th>
<th>Putative substrate</th>
<th>Putative substrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>McnA</td>
<td>1098</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Module 1</td>
<td></td>
<td>C, A&lt;sub&gt;Gln&lt;/sub&gt;, T</td>
<td>DAWQTGLIDK</td>
<td>Gln</td>
<td>Gln</td>
</tr>
<tr>
<td>McnB</td>
<td>1105/1101</td>
<td>C, A&lt;sub&gt;Thr&lt;/sub&gt;, T</td>
<td>DFWNIGMVHK</td>
<td>Thr</td>
<td>Thr</td>
</tr>
<tr>
<td>McnC</td>
<td>4721</td>
<td>C, A&lt;sub&gt;Thr&lt;/sub&gt;, T</td>
<td>DAVHFLGATVFK</td>
<td>Ile</td>
<td>Ile/Thr</td>
</tr>
<tr>
<td>Module 4</td>
<td></td>
<td>C, A&lt;sub&gt;Ahp&lt;/sub&gt;, T</td>
<td>DVENAGVVTK</td>
<td>Ahp precursor</td>
<td></td>
</tr>
<tr>
<td>Module 5</td>
<td></td>
<td>C, A&lt;sub&gt;Thr&lt;/sub&gt;, T</td>
<td>DAWTHAALK</td>
<td>Phe</td>
<td>Phe</td>
</tr>
<tr>
<td>Module 6</td>
<td></td>
<td>C, A&lt;sub&gt;Thr&lt;/sub&gt;, T</td>
<td>DASTIAGVCK</td>
<td>Tyr</td>
<td>Tyr</td>
</tr>
<tr>
<td>McnD</td>
<td>625</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>McnE</td>
<td>1418/1387</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Module 7</td>
<td></td>
<td>C, A&lt;sub&gt;Gln&lt;/sub&gt;, T, TE</td>
<td>DAQDLGVDK</td>
<td>Gln</td>
<td>Ile</td>
</tr>
</tbody>
</table>

*The domains encoded by the genes are indicated as follows: C, condensation; A, adenylation; T, thiolation (peptidyl carrier); M, N-methyltransferase; TE, thioesterase.
†The binding pocket sequences were identified by aligning the A-domains of Mcn with the Phe-activating A-domain of GrsA (Conti et al., 1997).
‡Substrate as predicted from the specificity-inferring code of adenylation domains (Challis et al., 2000; Stachelhaus et al., 1999). The Ahp precursor was deduced from extensive similarity with the Ahp precursor-activating domain described by Rouhiainen et al. (2000), including the identical binding pocket residues.
§Substrate as predicted by sequence alignments and phylogenetic analyses.
module 4 activates the precursor to the unusual amino acid 3-amino-6-hydroxy-2-piperidone (Ahp). The prediction is based on the high degree of similarity with the corresponding domain in the *Anabena* operon sequence, including identical binding pocket residues. Based on this, the peptide produced from the operon must be a cyanopeptolin, since Ahp is characteristic for this class of cyclic peptides (Namikoshi & Rinehart, 1996). The other invariant amino acid residue in all cyanopeptolins is Thr, which fits well with our prediction that the A-domain of module 2 activates this amino acid (Fig. 1b, Table 2). The lack of epimerization domains or a racemase gene in the gene cluster indicates that all amino acids incorporated into the peptide are in the L-configuration.

Since N-methylation has been shown to occur on the amino acid monomer prior to peptide bond formation (Billlich & Zocher, 1987; Weckwerth et al., 2000), the presence of an NMT-domain in module 6 indicates that the amino acid activated and condensed by this module becomes methylated.

The presence of a halogenase in the NRPS indicates halogenation of an amino acid residue, as shown for anabaenopeptilide (Rouhiainen et al., 2000) and for other cyanopeptolins when the amino acid in position R2 is Tyr (R positions according to Namikoshi & Rinehart, 1996). In the *mcn* gene cluster the ORF (*mcnD*) encoding the halogenase is located between modules 6 and 7, suggesting that the amino acid activated by module 6 is halogenated.

The C-domain at the N-terminal end of peptide synthetase module 1 indicates that the first amino acid residue in the peptide is acylated by a fatty acid as described for other peptides (Hoffmann et al., 2003; Konz et al., 1999).

The only A-domain in the cluster for which there is a clear disagreement between phylogenetic analyses and binding pocket signatures is A-domain 7 in McnE, for which phylogenetic analysis predicts Ile and the binding pocket signature suggests Gln. This discrepancy will be addressed further in the Discussion, but we will point out here that Ile is the amino acid activated by the corresponding A-domain in the *Anabena* gene cluster, while the main peptide detected in the present *Microcystis* strain contains Gln in this position.

Based on the operon organization and sequence, the predicted structure of the peptide synthesized by this operon from *Microcystis N-C 172/5* is as shown in Fig. 1(b). Since our initial screening had shown that this strain most likely produces a single peptide, the similarities between the peptide structure determined by chemical analyses and the structural information deduced from the operon were not unexpected and indicate that the peptide is synthesized by the enzymes encoded by this operon.

**Analysis of flanking regions**

To what extent do the immediate flanking regions of *mcn* contain similar or homologous ORFs compared to flanking regions of other described peptide synthetase operons? We sequenced about 1000 bp downstream and 4000 bp upstream of the putative cyanopeptolin synthetase to look for other genes commonly found in the clusters of characterized peptide synthetases. The upstream region contains two ORFs, *mcnF* and *mcnG*, transcribed in the opposite direction from the cyanopeptolin synthetase gene cluster. The downstream region did not display any ORFs with notable lengths.

The putative start codon (ATG) of *mcnF* (2070 bp) is located 557 bp upstream of the start codon of *mcnA*. *mcnF* is preceded by a putative promoter region with an *E. coli* σ70-like −10 element (TATAAT) 104 bp upstream of the ATG start codon. The corresponding −35 element is less conserved (TTAGCA). *mcnF* encodes a protein of 689 amino acids revealing extensive similarity to ABC transporters associated with cyanobacterial NRPS gene clusters: NosG (69% identity) and NcpC (66% identity) from *Nostoc* spp., McyH from the *mcy* gene clusters from *Anabena* strain 90 (69% identity), *P. agardhii* (63% identity) and *Microcystis aeruginosa* (62% identity); and Ndal from *N. spumigena* (66% identity). A detailed sequence analysis of McnF revealed the presence of five putative transmembrane helices at the N-terminal end between residues 49 and 334. The putative C-terminal ATP-hydrolysing domain contains three characteristic conserved motifs (Supplementary Fig. S3): Walker motifs A (GxxGxGKT/S) and B (hhhhDEAT) and the linker peptide (LSGGQQQR) (Walker et al., 1982). The EEA motif, characteristic of the ABC import systems (Dassa & Hofnung, 1985), was not present, indicating that the ABC transporter encoded by *mcnF* is an ABC exporter.

The translation start of the putative 1167 bp ORF (*mcnG*) lies 62 bp downstream of the TAG stop codon of *mcnF*. We observed no obvious promoter region in front of this gene, which encodes a 389 amino acid hypothetical protein with no obvious similarity to any proteins with known functions.

**Phylogeny of the different cyanopeptolin domains**

Phylogenetic analyses of the A-, T- and C- domains were performed separately at both the amino acid and nucleotide level (only protein trees are shown since the results were the same). We compared the Mcn domains from *Microcystis N-C 172/5* to domains in microcystin synthetase (Mcy) from *M. aeruginosa* (Tillet et al., 2000), anabaenopeptilide synthetase (Apd) from *Anabaena* strain 90 (Rouhiainen et al., 2000), nostopeptilide A synthetase (Nos) from *Nostoc* sp. GS224 (Hoffmann et al., 2003), nostocyclopeptide synthetase (Ncp) from *Nostoc* sp. ATCC 53789 (Becker et al., 2004) and nodularin synthetase (Nda) from *N. spumigena* (Moffitt & Neilan, 2004).

The A-domains from the different NRPSs revealed clustering largely according to the type of amino acid activated. In the tree in Fig. 3, A-domains from the same species/operon generally do not form clades. Intraspecies clades are formed...
only in the few instances where two A-domains within the same operon activate the same amino acid (such as NosA4-Leu and NosC1-Leu; ApdB3-Thr and ApdA2-Thr). Domains activating aromatic amino acids form a well-supported clade (PP 1.00; BS 96%). Interestingly, within this clade A-domains activating the same amino acid do not group together. The Ahp precursor domains form a distinct clade. Some notable exceptions to this general picture are also seen. The A-domain from McnE forms a well-supported clade with the A-domains activating Ile (from ApD, NcpB and NosA), despite having a binding site signature that suggests that this domain activates Gln (see above). Also, the A-domain in module 1 in McnC, with an Ile binding site signature, clusters basally to the Thr and Ile branches with low/moderate support.

The C-domains reveal a different topology (Fig. 4). Although domains condensing L-amino acids/halogenated N-methyl amino acids and D-amino acids form separate clades (for L/D-amino acid issues see Discussion), many C-domains tend to group according to the operon affiliation (i.e. the Nos, Apd and Ncp C-domains group together, respectively). There is also a tendency that some C-domains group according to their position in the operons. The N-terminal C-domains form a third separate clade.

Phylogenetic analyses of T-domains revealed a topology resembling the C-domains, with grouping by operon affiliation and positions in operon (Supplementary Fig. S4). Particularly, a grouping of T-domains according to neighbouring domains was seen: $T_E$ – situated upstream of an epimerization (E)-domain; $T_C$ – situated upstream of a C-domain; and $T_{TE}$ – situated upstream of a thioesterase (TE)-domain.

**DISCUSSION**

**Correlation between peptide and operon**

The screening for peptides in *Microcystis N-C 172/5* revealed a single peptide produced in significant amounts.
Furthermore, the structure we have determined fits very well with structural predictions that can be derived from the NRPS modules found in the operon we cloned from the same strain. In addition, PCR with consensus A-domain primers indicate that there are no further NRPS modules in this genome. Taking the results together it is likely that the halogenated peptide is produced by the mcn operon – or at least by an operon with almost identical structure. It should be emphasized, however, that a certain genetic link between peptide and operon can only be obtained by a knock-out of the operon in N-C 172/5, which has proven to be very difficult in this strain.

**Biosynthesis of a new cyanopeptolin**

Comparison of cypep-984 with other cyanopeptolins isolated from *Microcystis* spp. reveals several similarities, such as acylation of the side chain amino acid (Gln) and presence of an N-methylated aromatic amino acid in position R2 (ClMeTyr). One difference from other *Microcystis* cyanopeptolins concerns the C-terminal amino acid (position R1), which normally is either Ile or Val (Bister *et al.*, 2004), whereas in cypep-984 Gln is found in this position. Interestingly, the phylogenetic analysis of A-domains (Fig. 3) showed clustering of the corresponding A-domain in the gene cluster described here with those activating Ile in other peptide synthetases, while the binding site signature suggests that this domain activates Gln. If we accept that cypep-984 is a product of the gene cluster, this could indicate that this A-domain activated Ile in the common ancestor (as it still does in cyanopeptolin synthetases of other characterized cyanopeptolins of *Microcystis*) and has been altered to a Gln-activating A-domain in the peptide synthetase producing cypep-984. If so, the fact that the ‘binding pocket’ signature is identical to that of at least one other Gln-activating domain, LchAA (Yakimov *et al.*, 1998), while the rest of the sequence is ‘Ile-like’, may constitute an interesting example of convergent evolution.

All amino acids in known cyanopeptolins isolated from *Microcystis* spp. are in the L-configuration (Bister *et al.*, 2004). The chemical methods we have used to predict the peptide structure do not discriminate between L- and D-isomers, but *in silico* analysis of the mcn gene cluster suggested that this is also so in cypep-984, except that phylogenetic analysis showed that the C-domain in McnB clusters with those condensing D-amino acids (Fig. 4) – suggesting that the amino acid in the side chain (Gln) is in the D-configuration. Also the T-domain (McnA) upstream of this C-domain clusters with Tg-domains, even though there is no epimerization domain in the mcn gene cluster (Supplementary Fig. S4). This T-domain also has a conserved motif typical for Tg domains (GGDSL), rather than the GGHSL that is typical for TC-domains (Linne *et al.*, 2001). This unexpected juxtaposition of a Tg-domain and a D-amino acid condensing C-domain without an intervening E-domain might indicate that a deletion of an E-domain at the end of mcnA has occurred.

**The mcn operon of *Microcystis* has an overall structure similar to the Anabaena apd operon**

The cyanopeptolin synthetase from *Microcystis* N-C 172/5 is the second cyanopeptolin synthetase so far described at the gene level. Previously, Rouhiainen *et al.* (2000) characterized an anabaenopeptilide synthetase gene cluster (apd) from *Anabaena* strain 90. A comparison of the apd and mcn gene clusters reveals overall similarities, even though the two first modules in Mcn are encoded by separate ORFs. The general arrangement and transcriptional orientation are the same, and both clusters contain genes encoding halogenases. A flanking ABC transporter is found in both. Differences between mcn and apd include the lack in mcn of the two genes (apdE and F) encoding putative tailoring proteins, ApdE (a putative methyltransferase) and ApdF (a putative 3-oxoacyl reductase). In addition, mcnG, encoding a protein with unknown function, has no counterparts in the apd gene cluster.

The ABC transporter is located on opposite sides of the operon in *Anabaena* and *Microcystis*. Furthermore, in the mcn operon the ABC transporter gene is immediately flanking mcnA, while two ORFs (apd E and F) are located between the ABC transporter gene and apdD in *Anabaena*. Also, the *Microcystis* ABC transporter gene is transcribed in the opposite direction to the operon, while the *Anabaena* transporter gene is transcribed in the same direction as the operon. These findings indicate that the flanking regions, although containing some common elements, are structurally less conserved than the operons themselves.

**Ahp: an unusual amino acid present in all cyanopeptolins**

The pathway leading to formation of Ahp, the unusual amino acid at position 4 in the cyanopeptolin ring, is not yet resolved. Rouhiainen *et al.* (2000) have proposed that Gln may be the precursor of Ahp since the ‘binding-pocket’ sequence shows the highest similarity to those activating Gln. Our phylogenetic analyses of A-domains (see Fig. 3) show that the A-domains activating the Ahp precursor do not cluster together with Gln- or Glu-activating domains, but form an independent clade. Extrapolating from what is known from prokaryotic A-domains (see Challis *et al.*, 2000) and what this study has indicated, namely that A-domains cluster according to function, it may be that the Ahp precursor is not Gln/Glu (or at least not activated by an A-domain with a Gln/Glu binding site signature). The lack of phylogenetic congruence between the Ahp A-domains and the Glu A-domains from Mcy and Nda may, however, be because the latter ones most likely activate the D-form, and not the L-form, of Glu (Nishizawa *et al.*, 2001; Sielaff *et al.*, 2003; Tillett *et al.*, 2000). More work is needed to resolve this issue.

Rouhiainen *et al.* (2000) have suggested that both the methyltransferase domain of module 5 in the anabaenopeptilide synthetase and ApdF – a protein with similarity to
3-oxoacyl [acyl-carrier protein] reductases – may be involved in Ahp formation. In Mcn there is no methyltransferase present in module 5, and a gene with similarity to apdF was not found near the mcn gene cluster. Thus, none of the genes or domains within the mcn cluster seems to be involved in the formation of Ahp. It may be that genes involved in Ahp formation are located elsewhere in the genome.

**Halogenases and halogenated cyanopeptolins**

In many cases, the number, type and position of halogen groups in organic compounds are important for biological activity (see Nogami et al., 1990; Otsuka et al., 2004; Sancelme et al., 1994). Concerning cyanopeptolins, both chlorinated and non-chlorinated peptides have been described, e.g. the congeners anabaenopeptilide 90A (non-chlorinated) and anabaenopeptilide 90B (chlorinated) (Fujii et al., 1996). Rouhiainen et al. (2000) showed the presence of a halogenase gene in the gene cluster responsible for the synthesis of these two variants. In addition, we have recently sequenced a cyanopeptolin gene cluster from *P. agardhii* with high overall similarity to *mcn*, but without a halogenase, indicating that chlorination activity is not an enzymatic feature of all cyanopeptolin synthetases (T. B. Rouuge, T. Rohrlack, A. Tooming-Klunderud, T. Kristensen and K. S Jakobsen, unpublished). These findings indicate that halogenation is not essential from a functional perspective. Cyanopeptolins inhibit serine proteases, and in the two known structures of peptidase active sites with bound cyanopeptolins, the aromatic amino acid in position R2 (in one case Tyr, in the other chlorinated Tyr) is located in a hydrophobic pocket near the active site (Lee et al., 1994; Matern et al., 2003). Obviously, the presence or absence of a chlorine atom on the Tyr ring may influence this interaction and modulate the inhibiting activity of the peptide, so that the production of a set of variant peptides might lead to inhibition of a wider selection of proteases. All speculations regarding the effects of chlorination of this class of peptides are, however, hampered by the fact that we do not actually know the ‘real’ biological functions of these peptides. The existence of both halogenated and non-halogenated cyanopeptolins may itself indeed play a role in regulation of the biological activity, or point towards multiple functions for cyanopeptolins. An experimental setup including targeted knockout of the halogenase of Mcn is likely to be rewarding.

**The evolution of NRPS clusters**

From an overall perspective, the two cyanopeptolin operons (*Anabaena* and *Microcystis*) are quite similar when it comes to module order and architecture. Taking a closer look at the C- and T-domains, which tend to cluster according to gene affiliation or position in the operon, we note that the first modules (up to the C-domain in module 5) are less similar between the two operons than module 6, the halogenase and module 7 (see Figs 1 and 4 and supplementary data).

The phylogenetic analyses display different topologies for the A-domains compared to the C- and T-domains. Whatever the mechanisms may be for the difference between the A and C/T-domains, our findings indicate that the evolution seen for these two domain groups can to some extent be regarded to be independent.

The phylogenetic analyses do not show any clear evidence for genetic recombination between the two cyanopeptolin gene clusters. This is in contrast to the *mcy* genes, where data suggest that the *nda* cluster in *Nodularia* was formed by transposition of *mcy* from *Microcystis*, followed by deletion of two NRPS modules (Moffitt & Neilan, 2004). Likewise, there is no clear evidence for recombination between cyanopeptolin operons and other NRPS genes clusters such as *mcy*. Since cyanopeptolins and microcystins are likely to have co-existed for a long evolutionary period in the same lineage, this may indicate that the *mcn* and *mcy* classes of peptide synthetases have evolved independently. Consequently, this may be an indication of distinct and non-overlapping functions and strong functional constraints. Previously, recombination has been found to be prevalent in the microcystin class of NRPS (Kurmayer et al., 2004; Mikalsen et al., 2003; Moffitt & Neilan, 2004; Tanabe et al., 2004). The lack of evident recombination in the *mcn* operon may be attributed to lack of information about potential donor and acceptor sequences, since unlike Kurmayer et al. (2004) and Mikalsen et al. (2003) we have investigated only a single strain of the genus here. Furthermore, the *Anabaena* and *Microcystis* strains compared here for cyanopeptolin gene clusters may be too diverse from an evolutionary point of view, thus making *Anabaena/Microcystis* recombination less likely. This would be in agreement with the findings of Rantala et al. (2004) indicating that the *mcy* operons in *Anabaena* and *Microcystis* represent ancient evolutionary clades. In line with the predictions by Rudi et al. (1998) that recombination will increase between closely related strains (i.e. within the same genus), recently supported by whole-genome analyses of cyanobacteria (Zhayxybayeva et al., 2006), we expect that analysis of a panel of *mcn* genes from related *Microcystis* strains most likely would detect recombination between the strains.

**ACKNOWLEDGEMENTS**

We are grateful to Randi Skulberg, NIVA, for providing strain N-C 172/5. We thank Leo Rouhiainen, Kaarina Sivonen and Trine B. Rouuge for fruitful discussions. This work was in part supported by projects 157338/140 and 159822/V40 from the Norwegian Research Council to K. S. J.

**REFERENCES**


Edited by: K. Forchhammer