U.PORTO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
UNIVERSIDADE DO PORTO

Evolutionary genetics of the AMT domain of the microcystin gene cluster in various cyanobacteria genera

Samira Joussef Piña

Dissertação de Mestrado em Contaminação e Toxicologia Ambientais

2010

Samira Joussef Piña

Evolutionary genetics of the AMT domain of the microcystin gene cluster in various cyanobacteria genera

Dissertação de Candidatura ao grau de Mestre em Contaminação e Toxicologia Ambientais submetida ao Instituto de Ciências Biomédicas de Abel Salazar da Universidade do Porto.

Orientador – Doutor Agostinho Antunes
Categoria – Investigador
Afiliação – Centro Interdisciplinar de Investigação Marinha e Ambiental CIIMAR
Co-orientador – Professor Doutor Vitor Vasconcelos
Categoria – Professor catedrático
Afiliação – Faculdade de Ciências da Universidade do Porto

Abstract

The microcystins (MC) are the most common toxins in the toxic blooms of cyanobacterias, and are produced by a wide variety of cyanobacteria. The microcystins share a common structure, with high variability in two of the seven positions allowing the inclusion of different amino acids that generates several MC isoforms, while all the other positions are conserved. The enzyme complex responsible for their biosynthesis is encoded by the microcystin synthetase gene cluster *mcy*, which comprises ten genes *mcyA-J* (in *Microcystis*) that encode a giant enzyme complex comprising peptide synthetases (NRPSs), polyketide synthases (PKSs), and additional tailoring enzymes. The *mcyE* is a true natural hybrid that encodes a mixed protein with NRPS and PKS modules involved in the addition and activation of the Adda moiety (crucial for the toxicity) into the MC molecule. The McyE contain a catalytic domain that has homology with the glutamate-semialdehyde-aminotransferases, a class-III aminotransferase (pyridoxal 5'-phosphate-dependent family). Furthermore, this aminotransferase domain (AMT) is located at the boundary of the PKS and NRPS domain in the McyE synthetase and it is rare in thiotemplate assembly lines. The gene clusters are responsible by the production of many molecules that are important drugs or harmful toxins that are a public health risk. So, it is necessary to know the rules that govern the gene cluster evolution to understand and exploit their capabilities. Thus, the aims of this study were: to investigate the selective forces acting on the evolution of the AMT domain in several MC-producers species, as well as to investigate signs of recombination and mutation using state-of-the-art bioinformatic approaches. The analyses of the sequences retrieved from the GenBank for the AMT domain showed that the main force that acts on the AMT of McyE is the purifying selection. However, it has been found also regions with relaxation of the selective constraints with ω values between 0.2 - 0.6. This less conservative regions occurred at intervals interrupting the more conserved regions, which are related to the binding site of the cofactor pyridoxal-5'-phosphate. It has been found significant evidence for positive selection acting on one amino acid residue 88K that is not directly related to the amino acid incorporated into the microcystin molecule. This finding could indicate that some other functionality could be under selection for this site, related namely with changes in the catalytic efficiency or interactions between neighboring domains and modules. No recombination event was found in McyE, suggesting that point mutations are the main cause of genetic variation in the AMT domain of McyE for the cyanobacteria genera analyzed in this work. The phylogenetic relationships of the AMT domains retrieved in this work are in congruence with the results of Rantala et al., (2004), which suggest an ancient origin to the microcystin synthetase genes.

Resumo

As microcistinas (MC) são as toxinas mais comuns no blooms tóxicos de cianobactérias, sendo produzidas por uma grande variedade de cianobactérias. As microcistinas compartilham uma estrutura comum, possuindo uma elevada variabilidade em duas das sete posições, o que permite a inclusão de diferentes aminoácidos que dão origem a várias isoformas de MC, enquanto as outras posições são conservadas. O complexo enzimático responsável pela sua síntese é codificado pelo cluster de genes da microcistina sintetase, que compreende dez genes *mcyA-J* (em *Microcystis*). Este codifica um complexo enzimático gigante composto por peptídeos sintetases (NRPSs), policetídeo sintases (PKSs), e enzimas tayloring. O McyE é um verdadeiro híbrido natural que codifica uma proteína mista, com módulos NRPS e PKS envolvidos na adição e ativação do aminoácido Adda (crucial para a toxicidade) na molécula da MC. O McyE contém um domínio catalítico que possui homologia com as aminotransferasas-semialdeído-glutamato, uma aminotransferasa classe III (família dependente do piridoxal-5'-fosfato). Além disso, este domínio aminotransferase (AMT) está localizado no limite dos modulos PKS é NRPS na McyE, e é raro em linhas de montagem com tiotemplados. Os clusters de genes são responsáveis pela produção de muitas moléculas como drogas importantes ou toxinas prejudiciais que são um risco para a saúde pública. Assim, é importante caracterizar a evolução do cluster de genes para compreender e explorar as suas capacidades. Os objectivos deste estudo compreenderan: a investigação das forças selectivas que actuaram no domínio AMT em várias espécies produtores das MC, bem como investigar sinais de recombinação e mutação utilizando análises bioinformáticas avançadas. O análise das sequências do domínio AMT obtidas no GenBank revelou que a principal força que actua sobre o domínio AMT de McyE é a selecção purificadora. No entanto, detectaram-se também regiões com relaxamento das restrições selectivas, apresentando valores de ω entre 0,2-0,6. As regiões menos conservadas ocorreram em intervalos interrompendo as regiões mais conservadas, relacionadas com o local de ligação do cofactor. Foram encontrados evidências significativas de selecção positiva no resíduo de aminoácido 88K não directamente relacionado com o aminoácido incorporado na molécula de microcistina. Este resultado pode indicar que uma outra funcionalidade pode estar sob selecção neste resíduo, nomeadamente relacionada com mudanças na eficiência do catalisador ou interacções entre os domínios e módulos vizinhos. Não foi encontrado nenhum evento de recombinação no McyE, sugerindo que as mutações pontuais são a principal causa da variação genética no domínio AMT de McyE nos géneros de cianobactéria analisados neste trabalho. As relações filogenéticas dos domínios AMT obtidas neste trabalho estão em congruência com os resultados de

Rantala et al. (2004), o que sugere uma origem ancestral dos genes da microcistina sintetase.

General index

Index of figures

Index of tables

1. - Introduction

It is well known that the microorganisms produce a sophisticated array of secondary metabolites, as antibiotics and toxins, to inhibit or kill other organisms. Such feature is very important by itself since the interactions between secondary metabolites and microbial communities may be very effective in maintaining the biodiversity (Czárán et al., 2002).

Many of the current most important drugs are secondary metabolites, such as nonribosomal peptides (NRPs) and polyketides (PKs), produced by microorganism assembling big multifunctional enzyme systems called nonribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs), respectively. Typically the molecules are building from simple blocks such as carboxylic acids and amino acids (Minowa et al., 2007). There are two types of PKS, the type I and the type II. Each of the classes of PKSs resembles one of the classes of fatty acid synthases (FAS). The type I PKS possess a multidomain architecture similar to the type I FAS of fungi and animals, and type II PKS carry each catalytic site on a separate protein (Jenke-Kodama et al., 2005).

PKSs type I and NRPSs are very similar in the modular architecture of their catalytic domains and assembly-line mechanisms. Some peculiarities in the nonribosomal peptides (NRPs) like the presence of non-proteinogenic amino acids (as ornithine and dihydroxyphenyl-glycine), the presence of products like lipids or carbohydrates, and the great structural diversity (Schwarzer et al., 2003) make them very interesting. In fact, ever since the 1960's these molecules are targets of study, when experiments demonstrated that the synthesis of the NRPs was observed even in the presence of RNAses or inhibitors of the ribosomal biosynthesis (Gevers et al., 1968). Besides being toxic for microorganism and higher eukaryotes, the PKs may also be closely linked to microbial differentiation (Black and Wolk, 1994).

1.1. - An overview of NRPS and PKS

The NRPS and the PKS type I are megasynthetases structured in multimodular enzymatic assembly lines in which each module corresponds to one building block extension cycle, and the specific order of the modules defines the sequence of the amino acid (block) incorporated (Minowa et al., 2007; Wenzel and Müller, 2005). The modules can be further divided into domains which represent the enzymatic units that catalyze the individual steps of nonribosomal peptide synthesis and poliketide synthesis, which can be recognized at the protein level as highly conserved sequence motifs (Schwarzer et al., 2003). In the NRPS the elongation module needs at least three domains as the basic equipment: (1) an adenylation (A)-domain that selects the substrate amino acid and activates it as amino

acyl adenylate; (2) a peptidyl carrier protein (PCP)-domain that binds the co-factor 4′-phosphopantetheine (4′-PP) to which the activated amino acid is covalently attached; and (3) a condensation (C)-domain that catalyzes peptide bond formation (Schwarzer and Marahiel, 2001) (Fig.1).

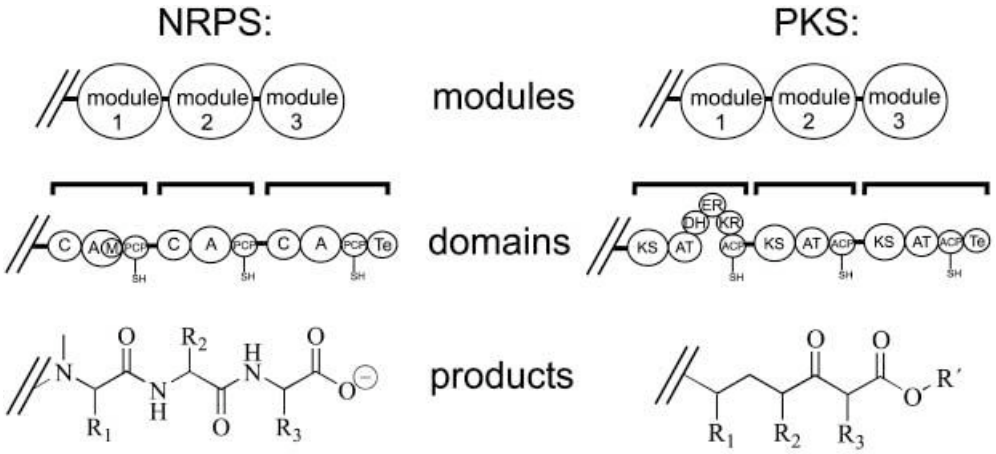

Fig.1. Simplified scheme from modules to products of the NRPS and PKS. In the NRPS: adenylation (A); peptidyl carrier protein (PCP); condensation (C); Thioesterase (Te). In the PKS: ketoacyl synthase (KS); acyltransferase (AT); acyl carrier protein (ACP); ketoreductase (KR); dehydratase (DH); enoyl reductase (ER); Thioesterase (Te) (Schwarzer and Marahiel, 2001).

Moreover, additional domains have been found in the NRPS modules, which expand the structural diversity of the synthesized peptides. So, it may be possible to obtain a billion of structures by using several molecular diversifying strategies like different combinations of the building blocks in the elongation steps, the degree of reduction after the condensation reactions, and the post synthetic processing of the products, such as cyclizations and glycosylations (Minowa et al., 2007). Curiously, structural analysis of free-standing condensation, adenylation, peptidyl carrier protein and thioesterase domains derived from NRPSs indicate that they are all monomeric, suggesting that the entire NRPS assembly lines may be constituted from “single-track” monomeric modules unlike PKS (Smith, 2002). In addition, the results of gel filtration, equilibrium ultracentrifugation and chemical crosslinking studies indicate clearly that NRPS modules are monomeric (Sieber et al., 2002). Also a good number of domains appear to act as independent catalytic units, being found on a single polypeptide chain. One example is the case of the adenylation domain, which is the core of each module, acting independently at the level of substrate recognition and activation (Marahiel et al., 1997).

On the other hand, the minimal module for operation of modular PKS type I is composed

of a ketoacyl synthase (KS) domain, an acyltransferase domain (AT), and an acyl carrier protein (ACP) domain. It is also frequently found the ketoreductase (KR), the dehydratase (DH), and the enoyl reductase (ER) domains embedded in the multifunctional megasynthases (Jenke-Kodama et al., 2005). In the PKS as well in the NRPS, the assembled molecule is ultimately released from the enzyme complex, usually by a thioesterase (TE) domain, which may also cause the direct cyclisation of the final product (Kalaitzis et al., 2009) (Fig.1).

The NRPS are also frequently found in association with PKS, which give rise to another variety of products with hybrid characteristics. This functionally compatible association is commonly called mixed NRPS/PKS system. There are several examples of these mixed assembly systems. However, there is a lack of knowledge about the evolution of these hybrids polyketide/peptide synthetases, and it is unclear whether the combination of these two systems is of recent origin. Some evidence demonstrates that the combination of these two systems is an ancient collaboration in the production of microcystins (Rantala et al., 2004).

There is a paradigm of how these protein systems work, i.e. each module is responsible for one elongation step and the specific order of the modules defines the sequence of the incorporated amino acid in the NRPSs, but there are also some exceptions to the rule. For the NRPSs it has been discovered that the modules from bacterial megasynthetases can operate iteratively, resulting in products with additions that make them more elongated or can be skipped during the biosynthetic process leading to shortened products (Wenzel and Müller, 2005).

The particular products of NRPS, PKS and mixed NRPS/PKS are currently of great interest among the research community, with a large amount of studies made over the last few years, especially *in silico* research. This field of study has focused its attention on the development of software for the comprehensive analysis of the NRPs, allowing even to predict compounds based on the constitution of the modules, and the modeling of 2D and 3D molecular structures (Ansari et al., 2004; Caboche et al., 2009; Li et al., 2009; Weber et al., 2009). Moreover, there are studies dedicated to clarify the structures and synthetic route of new several secondary metabolites, as well as to re-engineer natural products in order to increase, or alter, their biological activities and synthesized new products with pharmaceutical interest (Sieber and Marahiel, 2005; Zhao et al., 2008).

The cyanobacteria are among the group of organisms producers of bioactive secondary metabolites. A wide variety of cyanobacteria such as *Anabaena*, *Aphanizomenon*, *Cylindrospermopsis*, *Microcystis*, *Nodularia*, and *Planktothrix* produce several toxins, including anatoxin-a, saxitoxins, microcystins, nodularins, and cylindropsermopsins.

Among these, the cyanobacterium *Microcystis* is the most representative genus of toxic bloom forming cyanobacteria (Kaneko et al., 2007).

Several natural products encoded by NRPS/PKS and produced by cyanobacterial species have been described to date. Examples include the microcystin (*mcy*), the nodularin (*nda*), the aeruginosin (*aer*), the cyanopeptolin (*mcn*), the anatoxin A (*ana*), the nostopeptolide (*nos*) and many others, with peptide, polyketide or mixed biosynthetic origin (Table 1) (Kalaitzis et al., 2009).

Table 1. Some cyanobacteria-derived biosynthesis gene clusters*.

Gene cluster	Source organism	Size (kb)	Biosynthetic origin
Microcystin (*mcy*)	*Microcystis aeruginosa* PCC7806 *Planktothrix rubescens* NIVA CYA 98**	55 50	Polyketide/peptide Polyketide/peptide
Nodularin (*nda*)	*Nodularia. spumigena* NSOR10	48	Polyketide/peptide
Nostopeptolide (*nos*)	*Nostoc* sp. GSV224	40	Polyketide/peptide
Aeruginosin (*aer*)	*Planktothrix agardhii* CYA126/8 *Planktothrix rubescens* NIVA CYA 98**	34 30	Peptide
Anatoxin A (*ana*)	*Oscillatoria* PCC 6506	29	Polyketide
Cyanopeptolin (*mcn*)	*Microcystis* N–C 172/5 *Planktothrix rubescens* NIVA CYA 98**	30 35	Peptide Peptide

* Kalaitzis et al. (2009), ** Rounge et al. (2009).

Of the cyanotoxins mentioned above, microcystins (MC) are the most frequently found in the toxic blooms of cyanobacterias. The enzyme complex responsible for their biosynthesis is encoded by the microcystin synthetase gene cluster *mcy* (Tillet et al., 2000), which encode a giant enzyme complex comprising peptide synthetases, polyketide synthases (PKSs), and additional tailoring enzymes (Christiansen et al., 2003).

1.2. - Structure and characterization of the *mcy* gene cluster

Dittmann et al. (1997) identified putative peptide synthetase genes in the microcystin-producing strain *Microcystis aeruginosa* PCC7806 through gene disruption by homologous recombination. Several following studies have made a complete description of the *mcy* gene cluster and the biosynthesis pathway of a microcystin (Nishizawa et al., 1999; Nishizawa et al., 2000; Tillett el al., 2000). It has been shown that the modular nature of NRPS (Nonribosomal peptide synthetase) and PKS (Polyketide synthase) systems are well represented by the *mcy* cluster, because it contains several functional domains organized in a “mix-and-match” line, resulting in a functional synthetase (Tillett et al., 2000).

The cloning and sequencing of the *mcy* cluster have revealed a highly conserved set of multidomain proteins depicting the same basic reaction steps in the microcystin biosynthetic clusters from *Microcystis* (Chroococcales), *Planktothrix* (Oscillatoriales) and *Anabaena* (Nostocales) (Tillett et al., 2000; Christiansen et al., 2003; Rouhiainen et al., 2004). It has been observed differences in the clusters relatively to the disposition of the genes, the localization and orientation of the promoter regions, and the content of genes not directly related with the peptide assembly (Welker and von Döhren, 2006). Nevertheless, the basic modular organization of the biosynthetic NRPS/PKS systems is clearly preserved (Fig. 2).

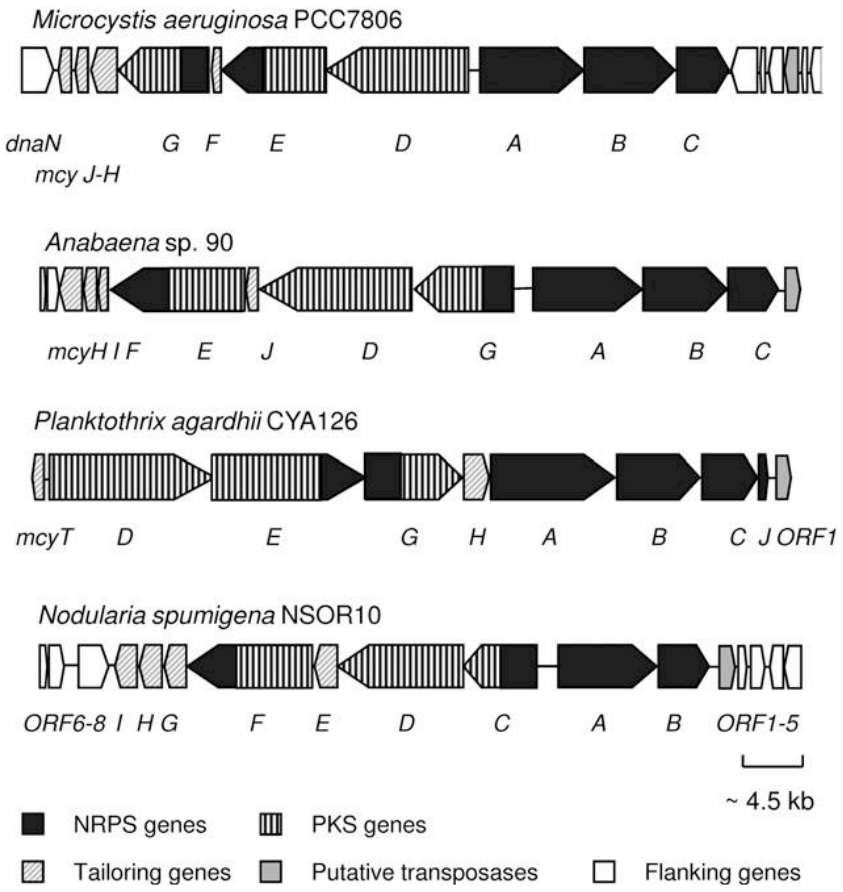

Fig.2. Microcystin biosynthesis clusters in *M. aeruginosa* PCC7806, *Anabaena* sp., *P. agardhii* CYA126 and *N. spumigena* (Kalaitzis et al., 2009).

The microcystins share the common structure cycle (Adda-D-Glu-Mdha-D Ala-L-X-D-MeAsp-L-Z), where X and Z are variable L-aminoacids, Adda is 3-amino-9-methoxy-2,6,8,-trimethyl-10-phenyldeca-4,6-dienoic acid, D-MeAsp is D-erythro-β-iso-aspartic acid,

and Mdha is N-methyl-dehydroalanine (Fig. 3) (Nishizawa et al., 1999). The positions X and Z show high variability allowing the inclusion of different amino acids, and all other positions are conserved. Thus, the possibility of different combination of the individual moieties allowing the variability of MC is appreciable. In fact, in the past years there have been described nearly 90 structural variants of MC (Welker and von Döhren, 2006).

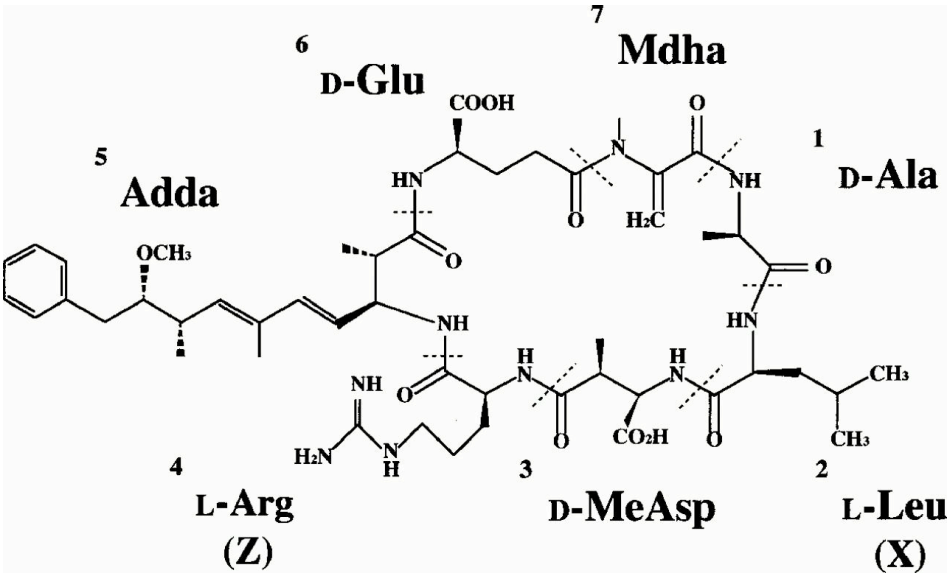

Fig.3. General structure of microcystins. The amino acids 2 (X) and 4 (Z) are variable. D-MeAsp: D-erythro-β-methylaspartic acid; Adda: (2S, 3S, 8S, 9S)-3-amino-9-methoxy-2,6,8-trimethyl-10-phenyldeca-4,6-dienoic acid; MdhA: N-methyldehydroalanine (Tillet et al., 2000).

Interestingly only a single *mcy* gene cluster is present in the genome of strains producing two or more structural variants of microcystins (Nishizawa et al., 1999; Tillett et al., 2000). Thus, because the adenylation domains (A-domain) of *mcy*A and *mcy*B are involved in the inclusion of amino acids at variable positions, it has been proposed that either the genetic variation in the A-domain of *mcy*B1 module of microcystin synthetasa genes (Mikalsen et al., 2003), or the nonspecific amino acid recognition by A-domains encoded by the *mcy* genes (Kurmayer et al., 2002), may be responsible for the variation of amino acids in the MC. Therefore these results suggest that to some extend the MC variation is genetically controlled (Tanabe et al., 2009).

1.3. - The *mcy* gene cluster in *Microcystis* as a model

In the analysis conducted by Tillett et al. (2000) it has been revealed six large open reading frames (ORFs) ranging from *mcyA* to *E* and *G* of a mixed NRPS/PKS nature,

together with an additional four small ORFs, *mcyF* and *H* to *J* (Fig. 2). The putative operon *mcyD-J* encodes the PKS/NRPS modules catalyzing the formation of the β-amino acid Adda and its linkage to D-glutamate, whereas *mcyA-C* encodes the NRPS modules for the extension and subsequent peptide cyclization. The *mcyD-J* operon is transcribed in opposite direction to the putative *mcyABC* operon. Both operons are transcribed from a central promoter between *mcyA* and *mcyD* (Kaebernick et al., 2002). The characteristics of the genes from each operon are summarized in the Table 2, and the predicted biosynthetic pathway is shown in the Fig.4.

Table 2. The characteristics of the microcystin *mcy* gene cluster in *Microcystis*, following Tillet et al., (2000).

ORF	Length (bp)	Putative RBS or translation initiation site	Encode (Da)	Function
***mcyD-J* operon**				
dnaN	422	569 bp downstream of *mcyJ*	-	DNA polymerase III beta subunit
mcyD	11721	733 bp upstream of *mcyA*	435714 polypeptide	Polyketide synthase type 1
mcyE	10464	167 bp downstream of stop codon of *mcyD*	392703 polypeptide	Polyketide synthase; peptide synthetase fusion protein
mcyF	756	32 bp downstream of stop codon of *mcyE*	28192 polypeptide	Amino acid racemase homolog
mcyG	7896	132 bp downstream of stop codon of *mcyF*	294266 polypeptide	Peptide synthetase; polyketide synthase fusion protein
mcyH	1617	224 bp downstream of stop codon of *mcyG*	67100 transmembrane protein	ABC transporter ATP-binding protein homolog
mcyI	1014	39 bp downstream of stop codon of *mcyH*	36838 polypeptide	D-3-phosphoglycerate dehydrogenase homolog; PGDH
mcyJ	837	176 bp downstream of stop codon of *mcyI*	31904 polypeptide	Putative O-methyltransferase
***mcyABC* operon**				
mcyA	8388	From the second ATG codon	315717 NRPS	Peptide synthetase
mcyB	6318	15 bp downstream of stop codon of *mcyA*	242334 PS	Peptide synthetase
mcyC	3876	4 bp downstream of stop codon of *mcyB*	147781 PS	Peptide synthetase
uma1	2051	-	-	Hypothetical protein
uma2	563	-	-	Hypothetical protein
uma3	1526	-	-	Hypothetical protein
uma4	1214	-	-	Transposase homolog
uma5	434	-	-	Hypothetical protein
uma6	970	-	-	Hypothetical protein

RBS: Ribosome binding site; PS: Peptide synthetase.

Additionally to these two operons, it has been found an *uma1-6* region with no function assigned. Interestingly, the *uma4* encodes a large peptide with similarity to TnpA, a transposase isolated from *Anabaena* sp. PCC1720 (Nishizawa et al., 2007). Besides, it has been found a region transcribed in the opposite orientation to the *mcyD-J* operon, localized at 569 bp downstream of the *mcyJ* and very similar to the DNA polymerase III β

subunit, called *dnaN*. This region is considerable conserved among the genus *Microcystis* (Nishizawa et al., 2007).

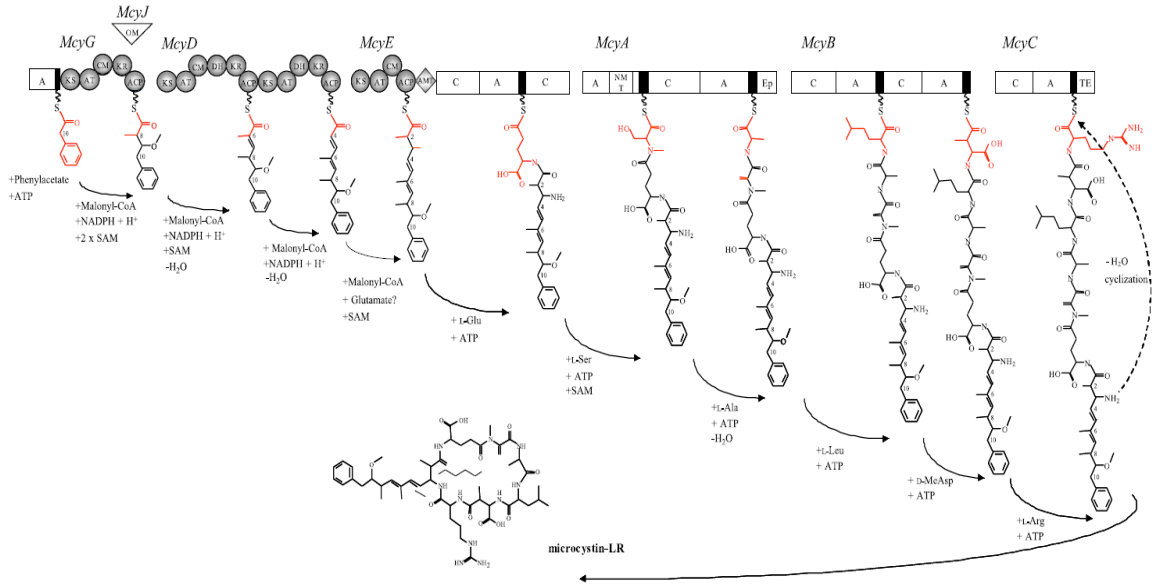

Fig.4. Model for the formation of microcystin and the domain structures of the six multienzymes McyA-E, G. Peptide synthetase domains: in white rectangles; Polyketide synthase domains: grey circles; KS: β-ketoacyl synthase; AT: acyltransferase; ACP: acyl carrier protein; KR: ketoacyl reductase; DH: dehydratase; CM: C-methyltransferase; OM: O-methyltransferase; A: aminoacyl adenylation; C: condensation; AMT: aminotransferase (Dittmann and Börner, 2005).

Kaneko et al. (2007) found numerous mobile elements among the entire genome of *Microcystis aeruginosa* NIES-843, which exhibited an exceptionally high content of mobile DNA elements. A total of 688 kb of the genome (11.8% of the entire genome) were occupied with insertion sequences (ISs) (583 kb, 10.0% of the genome) and miniature inverted-repeat transposable elements (MITEs) (105 kb, 1.8% of the genome). According also to previous studies some of these mobile elements were found in the proximities of the *mcy* gene cluster. So, it has been hypothesized that these elements are involved on the transference and functionality of the *mcy* gene cluster in the genus. Moreover, the presence of transposable elements in *Planktothrix* spp. has been related with the inactivation of the *mcy* gene cluster (Christiansen et al., 2006) and therefore the production of microcystins.

1.4. - Molecular regulation of the microcystin biosynthesis in *Microcystis*

The function of microcystins is still obscure, however, some experiences have been carried out to try to understand the regulation of its biosynthesis pathway. Also the increasing genome information available (Table 3) is allowing to improve the molecular

understanding of the biosynthesis of cyanotoxins. Several environmental factors have been described as possible influencing the biosynthesis of cyanotoxins. Parameters such as light intensity, temperature, nutrients and trace metals, have been tested under laboratory conditions and investigated with respect to their effect on cyanotoxin production (Kaebernick and Neilan, 2001).

Table 3. Currently sequenced cyanobacterial genomes in the GenBank.

Strain	**Genome size** (bp)	**Institution and submission date**
Acaryochloris marina MBIC11017	8 361 599	Translational Genomics Research Institute. 27 Aug 2007
Anabaena variabilis ATCC29413	7 068 601	US DOE Joint Genome Institute. 12 Sep 2005
Crocosphaera watsonii WH8501	6 238 156	US DOE Joint Genome Institute. 13 Jun 2005
Gloeobacter violaceus PCC7421	4 659 019	Kazusa DNA Research Institute, Chiba, Japan. 15 Aug 2003
Lyngbya aestuarii	7 087 904	CCY9616 J. Craig Venter Institute. 14 Dec 2006
Microcystis aeruginosa NIES-843	5 842 795	Kazusa DNA Research Institute, Chiba, Japan. 22 Nov 2007
Nodularia spumigena CCY9414	5 357 061	J. Craig Venter Institute. 25 Oct 2005
Nostoc punctiforme PCC73102	9 059 191	US DOE Joint Genome Institute. 07 Apr 2008
Nostoc sp. PCC7120	7 211 789	Kazusa DNA Research Institute, Chiba, Japan. 02 May 2001
Prochlorococcus marinus AS9601	1 669 886	J. Craig Venter Institute. 06 Nov 2006
Prochlorococcus marinus CCMP1375	1 751 080	Genoscope – Centre National de Sequencage, Evry, France. 28 May 2003
Prochlorococcus marinus MED4	1 657 990	DOE Joint Genome Institute. 03 Jul 2003
Prochlorococcus marinus MIT 9211	1 688 963	Massachusetts Institute of Technology. 24 Oct 2007
Prochlorococcus marinus MIT 9215	1 738 790	US DOE Joint Genome Institute. 07 Sep 2007
Prochlorococcus marinus MIT 9301	1 641 879	J. Craig Venter Institute. 16 Feb 2007
Prochlorococcus marinus MIT 9303	2 682 675	J. Craig Venter Institute. 10 Jan 2007
Prochlorococcus marinus MIT 9312	1 709 204	US DOE Joint Genome Institute. 27 Jul 2005
Prochlorococcus marinus MIT 9313	2 410 873	US DOE Joint Genome Institute. 03 Jul 2003
Prochlorococcus marinus MIT 9515	1 704 176	J. Craig Venter Institute. 06 Nov 2006
Prochlorococcus marinus NATL1A	1 864 731	J. Craig Venter Institute. 06 Nov 2006
Prochlorococcus marinus NATL2A	1 842 899	US DOE Joint Genome Institute. 08 Aug 2005
Synechococcus sp. CC9311	2 606 748	The Institute for Genomic Research. 04 Aug 2006
Synechococcus sp. CC9605	2 510 659	US DOE Joint Genome Institute. 27 Jul 2005
Synechococcus sp. CC9902	2 234 828	US DOE Joint Genome Institute. 08 Aug 2005
Synechococcus elongatus PCC6301	2 696 255	Center for Gene Research, Nagoya University, Japan. 10 Dec 2004
Synechococcus elongatus PCC7942	2 742 269	US DOE Joint Genome Institute. 08 Aug 2005
Synechococcus sp. RCC307	2 224 914	Genoscope – Centre National de Sequencage, Evry, France. 19 May 2006
Synechococcus sp. WH8102	2 434 428	US DOE Joint Genome Institute. 25 Jun 2001
Synechococcus sp. WH7803	2 366 980	Genoscope – Centre National de Sequencage, Evry, France. 19 May 2006
Synechococcus sp. JA-3-3Ab	2 932 766	The Institute for Genomic Research. 10 Mar 2006
Synechococcus sp. JA-2-3Ba	3 046 682	The Institute for Genomic Research. 21 Mar 2006
Synechocystis sp. PCC6803	3 947 019	Kazusa DNA Research Institute, Chiba, Japan. 01 Nov 2001
Thermosynechococcus elongatus BP-1	2 593 857	Kazusa DNA Research Institute, Chiba, Japan. 05 Jun 2002
Trichodesmium erythraeum IMS101	7 750 108	US DOE Joint Genome Institute. 21 Jun 2006

The transcriptional response of the microcystin biosynthetic gene cluster of *M. aeruginosa* PCC7806 was assessed by Kaebernick et al. (2000). In this study, *M. aeruginosa* PCC7806 was grown either under continuous light of various intensities, or under low light with subsequent short-term exposure to different light intensities and qualities, as well as various stress factors. The results of those experiences show that the light quality affects the microcystin synthetase production, either directly or via another regulatory factor, and require certain threshold intensities for their initiation, but possibly have a constitutive production. However, the results are not conclusive and the exact regulation process of the *mcy* gene cluster transcription and the release signaling remains unknown.

With regard to the regulation of release of microcystins, it was proposed that due to the lack of correlation between increased transcription and cellular toxin content, the toxin is release by a transporter (McyH) encoded within the *mcy* gene cluster. This transporter clustered phylogenetically with members of the ABC-A1 subgroup of the ABC ATPases (Pearson et al., 2004). In that study it was also found that a disruption of *mcy*H via deletional mutagenesis lead to the complete abolition of microcystin production under laboratory conditions, probably due to decrease of stability of the microcystin synthetase complex.

1.5. - Presence of the *mcy* gene cluster and toxicity

All genera with microcystin-producing strains also contain related strains that lack the ability to produce this toxin and the genera *Microcystis* is not an exception (Rantala et al., 2004).

It is currently know that the difference between microcystin-producing (toxic) and nonproducing (nontoxic) strains of cyanobacteria lies primarily in the presence or absence of microcystin synthetase gene cluster (Nishizawa et al., 2000). Furthermore, the results of the phylogenetic analysis made by Rantala et al. (2004) suggest that the microcystin synthetase genes were originally present, and that nontoxic strains of *Microcystis* have lost the ability to produce microcystins (derived character). They indicate that microcystin biosynthesis is a very ancient secondary metabolic pathway and there is no lateral transfer of *mcy* gene clusters between the genera. However, the relationship between toxigenicity and phylogeny within the cyanobacterial microcystin producers is unclear.

To clarify this issue several investigations have been made. Thus, the disruption of some *mcy* genes like *mcyA*, *mcyB* and *mcyH* have been performed, which demonstrated the important role of these genes in the microcystin biosynthesis pathway, causing the reduced expression in Δ*mcyA*, Δ*mcyB*, and Δ*mcyH* mutants (Pearson et al., 2004).

By contrast, Kaebernick et al. (2001) detected a strain that in fact contain the *mcy* genes but are unable to express them. This strain is a spontaneous mutant in microcystin

production (MRC) and it is genetically similar to a wild type (MRD), which produces a variety of rare microcystins. All the involved loci were examined, which allowed confirming the presence of the microcystin synthetase gene cluster for both MRD and MRC. Additionally, no mutations were detected in the sequences of the putative promoter regions for each of the *mcyA-J* genes, and also no toxin or toxic activity was detected for MRC. So, more studies are still necessary to understand the changes occurring between spontaneous mutant and wild type for microcystin production.

Furthermore, the lack of *mcy* genes are also described in other cyanobacteria genera like *Planktothrix*. Indeed, Christiansen et al. (2008) have found that all nontoxic strains analyzed have lost 90% of the microcystin synthetase (*mcy*) gene cluster, but those strains still contain the flanking regions of the *mcy* with remnants of transposable elements. The operon remnants and the flanking regions of those strains that lost the *mcy* gene cluster are represented in the Fig.5.

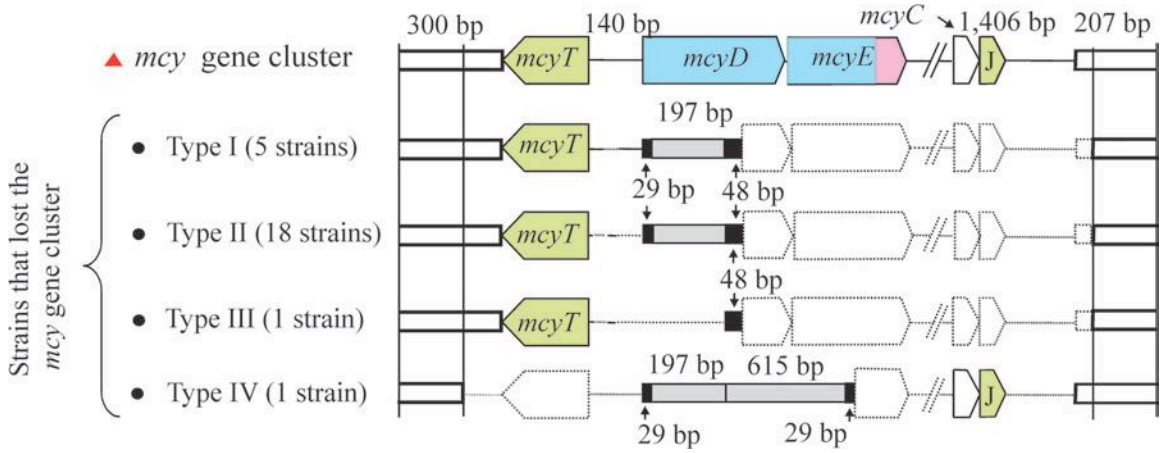

Fig.5. Representation of the *mcy* operon remnants and flanking regions in strains of *Planktothrix* that lost the *mcy* gene cluster (Christiansen et al., 2008). I–IV: types of gene cluster deletion events; Vertical straight lines: enclose the identical 5' and 3' ends; Gray gene regions: represent the remnants of the IS elements containing terminal inverted repeats (black boxes); Dotted areas: indicate the deletions; Red triangle: strains containing the *mcy* gene cluster; Black dots: nontoxic strains lacking the *mcy* gene cluster.

1.6. - Evolutionary history of the *mcy* gene cluster

The toxicity is wide spread through cyanobacterial genera, and toxic and non-toxic strains coexist in nature, so many studies have focused on unraveling the evolutionary history of microcystins and why the presence of microcystins have a patchy distribution. To try to explain why so distant genera share the *mcy* gene cluster it has been proposed i) that the production of microcystin has an ancient origin from a common ancestral cyanobacterium,

thus the heterogeneous distribution of the toxic and non-toxic strains may be the result of gene deletions at a number of times throughout evolution (Rantala et al., 2004); or ii) that the microcystin production has a recent origin by horizontal gene transfer (HGT) and recombination events (Otsuka et al., 1999; Tillet et al., 2001). These hypotheses remain in debate.

The results obtained by Rantala et al. (2004) highlight this issue, since in their work it was found evidence for the early evolution of microcystin synthesis and no evidence for lateral transfer of the *mcy* gene clusters between the genera. This has been support by other studies, which have suggested a monophyletic origin of the *mcy* gene cluster (Christiansen et al., 2008). In addition, Rantala et al. (2004) also proposed a recent origin of nodularins from microcystins, which is consistent with the production of nodularins by a single cyanobacterial genus and the limited structural variation of nodularins in comparison to microcystins, but they do not excluded the possibility of a more recent origin of parts of the microcystin synthetase gene cluster between strains within a genus by HGT.

On the other hand, it has been suggested that recombination is an important factor on the genetic variation in bacteria. Tanabe et al. (2004) found evidence that intragenic and intergenic recombination contributes to the genetic diversity of the *mcy* gene of different strain of *Microcystis* spp., thus propose that the evolutionary history of the *Microcystis* genes is reticulate due to the mosaic structure of the *mcy* gene cluster. But they also speculated that the presence of undiscovered plasmids could possibly mediate the exchange of genetic material between *Microcystis* strains, reducing the frequency of recombination between *mcy* genes, and thus becoming more incidental in the evolutionary history of microcroystis.

Also, Fewer et al. (2007) found recurrent adenylation domain replacements by recombination without the transference of the condensation domains and showed that the *mcyB1* and *mcyC* adenylation domains (which recognize and activate the amino acids found at X and Z positions) are recombination hotspots in the microcystin synthetase gene cluster.

Tanabe et al., (2007) used a multilocus sequence typing (MLST) comprehending housekeeping loci and *mcy* genes to investigate the evolutionary genetic background of *Microcystis aeruginosa*. Their results point to a later acquisition of the toxic genes in the microcystin-producing strains, providing further evidence for the gain and loss of toxicity during the intraspecific diversification of *M. aeruginosa*.

More recently, Tanabe et al. (2009) found that individual phylogenetic and multilocus genealogical analysis of the *mcy* genes, divided their isolates of toxic and nontoxic strains of *M. aeruginosa* into two clades, consistent with previous phylogenetic results obtained

by MLST, based on seven housekeeping loci (Tanabe et al., 2007). Their population genetic data corroborate previous findings about the importance of recombination. In fact, population recombination rates were larger than mutation rates, suggesting that the frequency of recombination is greater than point mutation. Such study also indicated that recombination seems to be higher within groups than between them, implying that there exists genetic isolation between groups, also evident by the strong genetic clustering of *mcy* genes. Finally, they concluded that recombination and neutral genetic drift are primarily responsible for the observed deep divergence of the *mcy* genes in *M. aeruginosa*.

1.7. - The McyE gene as a key gene in the production of MC and in evolutionary studies

Several studies have evaluated the genetic variability of the genes involved in the production of microcystins isoforms (Kurmayer and Gumpenberger, 2006; Fewer et al., 2007; Tooming-Klunderud et al., 2008). By contrast, some works have been performed using genes of the MC cluster that are less variable and are not involved in the production of isoforms. Genes like McyE, *G* and *D* that are involved in the addition and activation of the Adda moiety into the MC molecule are currently used as molecular markers (Rantala et al., 2006; Mbedi et al., 2005; Vasconcelos et al., 2010) to evaluate the presence of microcystin producer species in water bodies by using specific primers for these genes, which is an important issue for the public health.

Within the MC genes, the McyE gene encodes the hybrid PKS-NRPS protein McyE that have catalytic domains of both kinds of proteins. McyE is also a link between the PKS and the NRPS domains of the MC megasynthetases system (Fig.4). It is located at the end of the PKS assembly line and catalyzes the incorporation of D-Glu to the growing chain of the MC (Tillet et al., 2000). Furthermore, the McyE contain a catalytic domain that are rare in thiotemplate assembly lines (Aron et al., 2005), which is an aminotransferase domain (Fig.4) that is located at the boundary of the PKS and NRPS domain in the McyE synthetase. This domain is supposed to be involved in the transfer of an amino group to the side chain of the Adda moiety (Christiansen et al., 2003).

This type of proteins are called fusion proteins because conjugate the products of the subclusters and their evolutionary origins are particularly significant. An enzyme that performs the same function is also present in the nodularin gene cluster and is encoded by the *ndaF* gene (Jungblut and Neilan, 2006). The similarities between microcystins and nodularins have been evaluated previously and both toxins seem to have a common phylogenetic origin (Rantala et al., 2004).

The nature of the docking enzymes on hybrid multienzymes systems of secondary

metabolites composed by PKS and NRPS, and how the interdomain regions interact has been of particular interest (Broadhurst et al., 2003; Richter et al., 2007) due to their potential to be used in combinatory biosynthesis of natural products (Du and Shen, 2001; Du et al., 2003; Schwarzer and Marahiel, 2001; Aron et al., 2005). The subcluster docking zones provide a particularly interesting form of chemical innovation in which the products of a metabolic pathway merge, enabling new forms of structural diversity among small molecules (Fischbach et al., 2008).

The gene clusters are responsible by the production of many molecules that are important drugs or harmful toxins that are a public health risk. So, it is necessary to know the rules that govern the gene cluster evolution to understand and exploit their capabilities. Moreover, the molecular mechanisms involved in the synthesis of secondary metabolites are among life's most diverse and rapidly evolving genetic elements, which evolve over shorter time scales relatively to the genes from higher organisms (Fischbach et al., 2008), therefore being interesting elements for the study of molecular evolution.

1.8. - Objectives

Due to the particular importance of the *mcyE* in the assembly of microcystin and of the *ndaF* in the assembly of nodularins (given that the NdaF performs the same function that McyE and have a common phylogenetic origin), as well as the special characteristic of the aminotransferase catalytic domain described above, the aims of this study included:

- The molecular assessment of the AMT domains from the McyE and the NdaF genes in strains of microcystin and nodularin cyanobacteria producers genera, which include: *Microcystis*, *Anabaena*, *Planktothrix*, *Oscillatoria*, *Phormidium*, *Nostoc* and *Nodularia*.
- The determination of the role of recombination and mutation on the AMT domain of *mcyE* among several microcystin and nodularin producer strains.
- The assessment of the influence of Darwinian selection acting on the evolution of the AMT domain of *mcyE* among several microcystin and nodularin producer strains.
- The evaluation of the mechanisms related with the origin and evolution of the microcystins.

2. - Methods

2.1. - Sequences alignment

The nucleotide sequences from the Aminotransferase domain (AMT) of the McyE and the NdaF genes used in this study were retrieved from the GenBank (NCBI). All the sequences available in the GenBank for the AMT domain from the *mcy* gen cluster were collected. A total of 54 sequences were initially retrieved, but the removal of identical sequences and sequences with internal stop codons resulted in a final dataset of 20 sequences: 7 from *Microcystis*, 4 from *Anabaena*, 3 from *Phormidium*, 2 from *Planktothrix*, 2 from *Oscillatoria*, 1 from *Nostoc*, and 1 from *Nodularia*. Accession numbers are available in the Table 4.

The terminal stop codon of the sequences used in this study was removed and the sequences were aligned according to their translated amino acids in a multiple alignment with Clustal W v.2.0 (Larkin et al., 2007). The alignment was then inspected by eye followed by manual edition.

2.2. - Recombination analyses and nucleotide substitution statistics

The detection of potential recombinant sequences was investigated with RDP3 Beta 42 software (Martin et al, 2005a), to identify the most likely parental sequence and the localization of possible recombination breakpoints. The aligned sequences were examined using RDP (Martin and Rybicki, 2000), GENECONV (Padidam et al., 1999), BOOTSCAN (Martin et al., 2005b), MAXIMUM CHI SQUARE (Smith, 1992), CHIMAERA (Posada and Crandall, 2001) and SISCAN (Gibbs et al., 2000) in the RDP3. The default settings was used for all methods, were the sequences were considered as linear. The recombination was also evaluated using the SplitsTree v.4 software (Huson and Bryant, 2006) allowing the estimation of the recombination Phi test (Φ) (Bruen et al., 2006).

2.3. - Phylogenetic analyses

The phylogenetic analyses required the prior estimation of the best model of nucleotide substitution using Modeltest v. 3.7 (Posada and Crandall, 1998) and the empirical model of amino acid substitution using ProtTest v.2.4 (Abascal et al., 2005). The Neighbor joining method (NJ), the Maximum likelihood (ML) method and the Bayesian analysis were used to infer the phylogeny. The NJ tree was made under the model of maximum composite likelihood and using the default settings for NJ bootstrap analysis in MEGA 4 (Tamura et al., 2007). The ML tree was estimated using SeaView v.4 (Gouy et al., 2010) driving the program PhyML (Guindon and Gascuel, 2003) under the GTR model of DNA substitution determined by Modeltest. A 1000 bootstrap replicates was used in both NJ and ML

analysis. The Bayesian tree was obtained using the program Mr. Bayes v3.1 for Bayesian inference of phylogeny (Ronquist and Huelsenbeck, 2003), under the following parameters: the GTR model of DNA substitution with a gamma distribution of rates, running 2 million generations and sampling trees every 100 generations, burn-in 3000 trees. The software FigTree v.1.3.1 was used for visualizing the Bayesian tree. The trees topology obtained from all methods were congruent between them.

Table 4. Characteristics of the *mcyE* and the *ndaF* gene sequences of microcystin and nodularin producers genera used in this study.

Genera	**Accession number**	**Specie**	**Strain**	**bp**	**Hepatotoxin**	**Origin**	**Reference**
Microcystis	AY817159	*viridis*	NIES-102	417	Microcystin	Japan	Jungblut and Neilan, 2006.
	AY817158	*wesenbergii*	NIES-107	417	Microcystin	Japan	Jungblut and Neilan, 2006.
	AY817161	*aeruginosa*	UTEX LB 2664	417	Microcystin	United States	Jungblut and Neilan, 2006.
	AB032549	*aeruginosa*	K-139	10464	Microcystin	-	Nishizawa et al., 1999
	AY817160	*aeruginosa*	PCC 7005	417	Microcystin	Scotland	Jungblut and Neilan, 2006.
	AM778952	*aeruginosa*	PCC 7806	10467	Microcystin	-	Frangeul, Unpublished
	AY817162	*aeruginosa*	UTEX B 2667	417	Microcystin	United States	Jungblut and Neilan, 2006.
Oscillatoria	AY817165	sp.	19R	414	Microcystin	-	Jungblut and Neilan, 2006.
	AY817164	sp.	18R	414	Microcystin	Finland	Jungblut and Neilan, 2006.
Planktothrix	GQ451434	sp.	VUW25	335	Microcystin	New Zealand	Wood et a., 2010
	GQ451433	sp	CYN61	335	Microcystin	New Zealand	Wood et a., 2010
Phormidium	AY817166	sp.	2-26b3	417	Microcystin	United States	Jungblut and Neilan, 2006.
	AY817167	sp.	1-6c	417	Microcystin	United States	Jungblut and Neilan, 2006.
	AY817168	sp.	4-19b	420	Microcystin	United States	Jungblut and Neilan, 2006.
Nodularia	AY210783	*spumigena*	NSOR10	10428	Nodularin	-	Moffitt and Neilan, 2004
Nostoc	AY817163	sp.	152	417	Microcystin	Japan	Jungblut and Neilan, 2006
Anabaena	EU916757	sp.	315	418	Microcystin	Finland	Fewer et al., 2009
	EU916774	lemmermannii	PH256	418	Microcystin	Finland	Fewer et al., 2009
	EU916770	flos-aquae	NIVA-CYA 267/4	418	Microcystin	Finland	Fewer et al., 2009
	AY817157	sp.	202	417	Microcystin	Finland	Jungblut and Neilan, 2006.

2.4. - Selection analyses

The CODEML from PAML v4 package (Yang, 1997) was used to test the presence of codon sites affected by positive selection and to identify sites under selection using the likelihood approach. For these analysis it was employed the site models, and performed two likelihood-ratio tests, a test of M1a (nearly neutral) versus M2a (positive selection), and that of M7 (beta; assuming a beta distribution of dN/dS over sites ranging from 0 to 1) versus M8 (beta & ω; the same as M7 with an additional estimate of ω = dN/dS > 1) to determine any sites under positive selection (dN/dS > 1). The results were taken from the Bayes Empirical Bayes (BEB) (Anisimova et al., 2001, 2002; Yang et al., 2005) and Naive Empirical Bayes (NEB) output from each model. The M0 model was used to estimate the overall ω of the protein and for compare with the discrete model M3. A likelihood ratio test (LRT) (Yang et al., 2000) was estimated by comparing a model for selection (alternative model) with the model that does not account for selection (the null model).

Additionally, an analysis to determinate which codon sites are under diversifying, positive or negative selection was made using the Datamonkey webserver (http://www.datamonkey.org/) (Kosakovsky and Frost, 2005a; Kosakovsky et al., 2005). Three different codon-based maximum likelihood methods were used in Datamonkey for estimate the dN/dS (ω) ratio: Single likelihood ancestor counting (SLAC), Fixed effects likelihood (FEL) and Random effects likelihood (REL) (Kosakovsky and Frost, 2005b). All these analyses were run with the default options. SLAC infers selection by comparing the observed rates of non-synonymous and synonymous mutation at each codon to that expected under a binomial distribution. FEL compares the model fit in which non-synonymous and synonymous mutations are constrained to be equal, to an unconstrained model. REL methods approximate the distribution of non-synonymous to synonymous rates across all sites into classes, and calculate the posterior probability that each site belongs to each of the rate classes (Poon et al., 2009).

Furthermore, it was performed another analysis of the rate of synonymous substitutions (the number of synonymous substitutions per synonymous site, dS) and non-synonymous substitutions (the number of non-synonymous substitutions per non-synonymous site, dN), as well as their ratios (dN/dS) in a sliding window using the SWAAP v.1.0.3 software (Pride, 2000). The Nei–Gojobori distance estimation method was used for a sliding window (Nei and Gojobori, 1986). Also, it was performed a sliding window analysis of transitions and transversions.

Finally, the software ConTest (Constraint Testing) (http://home.gna.org/contest/) was used to detect amino acids with evolutionary constraint but accounting for the evolution of specific biochemical properties as volume, polarity and charge (Dutheil, 2008) under the James, Taylor and Thornton model (JTT) (Jones et al., 1992). To support statistically the

results, the analysis was conducted with the R statistical software (R Proyect) using the ade4 package (Chessel et al. 2004).

2.5. - Secondary structure of AMT domain of McyE

The online server I-Tasser (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) (Zhang, 2008, 2009; Roy et al., 2010) was used to made a prediction of the secondary structure and to obtain a 3D model of the protein structure of the AMT domain. This program also predicts the biological function of the protein by comparisons with the protein function database. The predicted binding sites (from I-Tasser) and the sites under positive selection were indicated in the output model using the software Discovery Studio Visualizer v.2 (http://accelrys.com/products/discovery-studio/visualization-download.php) from Accelrys Software Inc.

3. - Results

In this study the AMT domains of the microcystin synthetase McyE module and the nodularin syntethase NdaF module from 20 cyanobacteria strains have been compared. There are seven microcystin and nodularin producers species including *Microcystis*, *Anabaena*, *Planktothrix*, *Oscillatoria*, *Phormidium*, *Nostoc* and *Nodularia* (Table 4). The sequences alignment shown in Fig. 6 was used to perform the genetic analyses and the amino acids positions labeled in the alignment are used as reference positions along this study.

From the *mcy* gene cluster, it has been chosen for in deep study the *mcyE* because this gene encodes a mixed polyketide peptide synthetase (McyE) involved in the synthesis of the amino acid Adda and the activation and addition of D-glutamate into the MC molecule (Nishizawa et al., 2000; Tillett et al., 2000). These two amino acids are crucial for the microcystin toxicity and do not vary among the microcystins isoforms. Furthermore, this gene region has been successfully used as a reliable molecular marker for the detection of MC producers in several works (Rantala et al., 2006; Mbedi et al., 2005; Vasconcelos et al., 2010).

3.1. - Phylogenetic analyses based on the McyE aminotransferase domain

Phylogenetic analyses of the 20 McyE AMT domain sequences yielded a similar tree topology for all methods used (NJ, ML and Bayesian) with high bootstrap support at all major branches. Five well-supported clades were observed and perfectly separated the McyE by genera, reflecting a taxonomic feature that included the presence or the absence of heterocyst (Fig.7) represented by two different clusters (named as I and II), respectively. The cluster I is represented by the strains of the genera *Anabaena*, *Nostoc* and *Nodularia* belonging to the family Nostocaceae (with heterocyst), whereas cluster II contained the genera *Oscillatoria*, *Planktothrix*, *Phormidium* and *Microcystis* (without heterocyst). This result is consistent with the obtained by Jungblut and Neilan (2006). They also found that sequences from bloom samples of geographically close sites cluster together, having a high identity percent, even among the isolates of different species. In this work it was found a similar pattern, suggesting the importance of performing further analysis with a larger group of geographically diverse sequences to establish whether or not any phylogeographical pattern exist, because it is probable that the strains have diverged in other phenotypic characters that are not directly linked to the MC production.

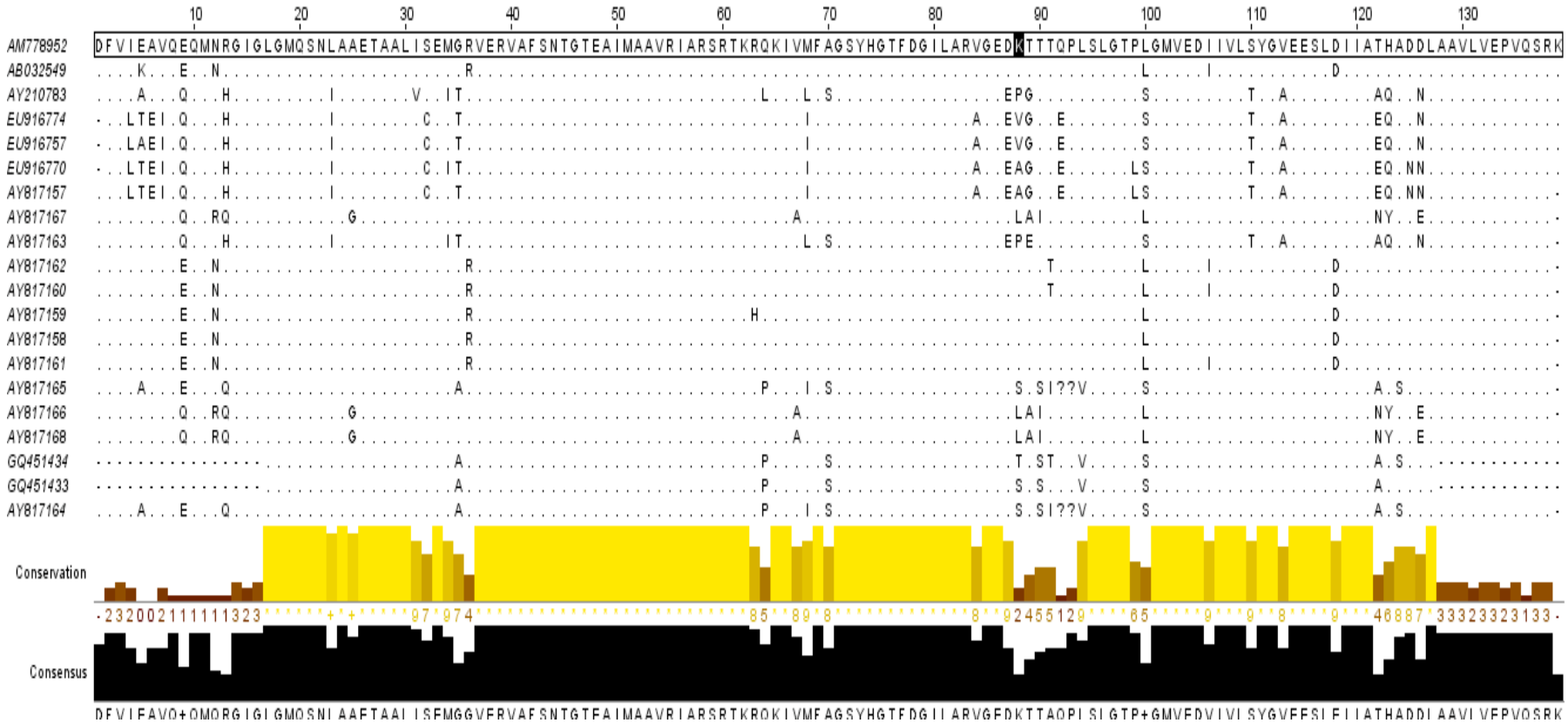
10 20 30 40 50 60 70 80 90 100 110 120 130
AM778952 DFVIEAVQEQMNRGIGLGMQSNLAAETAALISEMGRVERVAFSNTGTEAIMAAVRIARSRTKRQKIVMFAGSYHGTFDGILARVGEDKTTTQPLSLGTPLGMVEDIIVLSYGVEESLDIIATHADDLAAVLVEPVQSRK
AB032549
AY210783
EU916774
EU916757
EU916770
AY817157
AY817167
AY817163
AY817162
AY817160
AY817159
AY817158
AY817161
AY817165
AY817166
AY817168
GQ451434
GQ451433
AY817164
Conservation
Consensus

Fig.6. Alignment of the amino acid sequences of the AMT domain. The position 88 is highlighted with a black square. Histograms of the site conservation and the consensus sequence are shown below the alignment.

It has been found that the distribution of microorganisms is not restricted by geographical barriers (Finlay, 2002), as well as that the natural populations of bacteria and cyanobacteria have high recombination and migration rates that are not influenced by geographical factors (Roberts and Cohan, 1995; Barker et al., 2000; Lodders et al., 2005). Furthermore, in a study performed by Tanabe et al. (2009) there was a little geographic contribution to the pattern of genetic variation within the *mcy* genes in the microcystin producer genera. Although the effects of spatial separation cannot a priori be excluded, isolation is essential to understand the evolution of MC synthesis.

Since in this study it was included one sequence from the nodularin aminotransferase domain of NdaF (AY210783), it was also evaluated it relationship with the microcystin aminotransferase domain of McyE. Phylogenetic analysis indicated that the AMT domain of the nodularin sinthetase gene cluster is closer to the AMT domain of the microcystin sinthetase gene cluster from *Nostoc*, clustering together in a sub-branch of the phylogenetic tree (Fig.7). The species *Nostoc*, *Nodularia* and *Anabaena* are all heterocyst-forming. This result is consistent with the obtained by Jungblut and Neilan (2006) and Rantala et al. (2004). Furthermore, Moffitt and Neilan (2004) found that McyE and NdaF share 74% identical protein sequences, and both proteins have the same order of catalytic domains.

3.2. - Recombination analyses

Recombination is thought to be one of the main mechanisms driving the diversification of PKS and NRPS, and particularly the microcystin gene cluster (Tanabe et al., 2004; Fewer et al., 2007; Tanabe et al., 2009; Jenke-Kodama and Dittmann, 2009). So, the analysis of the nucleotide sequences of *mcyE* included the evaluation of recombination breakpoints in the AMT domain. Seven different methods to detect recombination (RDP, GENECONV, BOOTSCAN, MAXIMUM CHI SQUARE, CHIMAERA, SISCAN) have been used in the RDP3 software, and the Phi test (Φ) for recombination (a robust statistical test for recombination) was estimated in the SplitsTree software. These analyses found one signal of recombination supported only by two of the six methods, so the recombinant sequence was removed from the dataset. The reminder sequences did not show any signal of recombination.

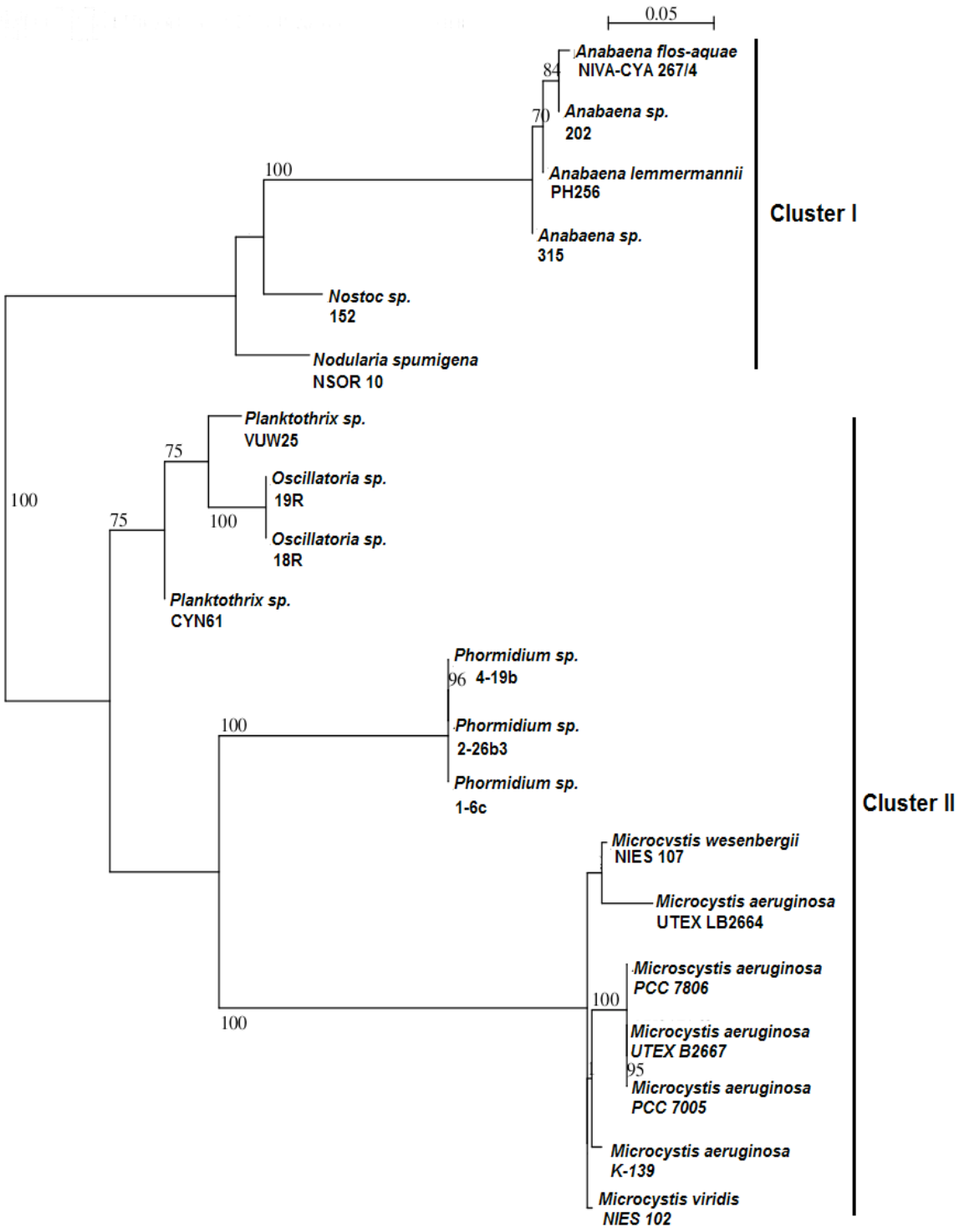

Fig.7. Phylogenetic tree based on ML analysis from the AMT domain translated amino acids sequences of McyE and NadF. Cluster I: cyanobacteria with heterocyst; Cluster II: cyanobacteria without heterocyst. Bootstrap percentages greater than 50 (after 1000 data resampling events) are shown.

Rantala et al. (2004) found congruence between genes involved in primary metabolism (16s rRNA and rpoC1) and genes involved directly in the synthesis of microcystins and nodularins (*mcyA*, *mcyD*, and *mcyE*), i.e. they found that no recombination event influenced significantly the phylogenetic relationship between their dataset and if any recombination event occurred must have been relatively recent. Also, Mbedi et al. (2005) found by comparing the nucleotide sequences of *mcyE* and *mcyB* within the genus *Planktothrix* that these are differently conserved, namely *mcyE* show less variability compared to the *mcyB*. Thus, the low variability of *mcyE* could be explained by the functional importance of McyE in the microcystin-biosynthesis. Indeed, the modules of McyE, together with McyD and McyG, are responsible for the synthesis of the unique Adda moiety of microcystins (Nishizawa et al., 2000; Tillett et al., 2000), an amino acid common to all MC isoforms and a key to the toxicity.

In addition, Tanabe et al. (2004) reported the existence of recombination within *mcyA*, but not within *mcyD*, *mcyG* and *mcyJ* genes of closely related *Microcystis* strains. However, it also was determined that incongruence between the four *mcy* gene phylogenies indicated that recombination should have occurred over the entire *mcy* gene cluster. Nevertheless, no sign of recombination has been detect previously or in this study within the AMT domain of *mcyE* in any cyanobacteria species.

3.3. - Selection analyses

It has been proposed that synonymous mutations are mostly invisible to natural selection, while non-synonymous mutations can be under strong selective pressure (Yang et al., 2000). Thus, comparison of the fixation rates of these two types of mutations provides a powerful tool for understanding how natural selection affects the molecular sequence evolution (Kimura, 1983). The non-synonymous/synonymous rate ratio ($\omega = dN/dS$) is an important indicator of selective pressure at the protein level, with $\omega = 1$ meaning neutral mutations, $\omega < 1$ purifying selection, and $\omega > 1$ diversifying positive selection.

A ML analysis was used to estimate the synonymous and non-synonymous substitution rates using the PAML software. Because this program exclude ambiguity sites (gaps) within the alignment, some sites were excluded and therefore the numbering of amino acids positions located after the ambiguity sites are changed, as well as the total number of amino acids positions analyzed by PAML, i.e. the initial and the final length of the alignment is 139 and 109 amino acids, respectively. However, for better understanding, the numbering of amino acids position assigned by the alignment is maintained across this study.

Thus, the total number of amino acids sites analyzed for estimate the synonymous and

non-synonymous substitutions was 109. The dN/dS (ω) ratio across all the sites was 0.076 (measured by model M0 in PAML). This low value indicates an excess of synonymous mutations, that is not surprising because many coding sites experience purifying selection or are largely invariable (with ω close to 0) due to strong functional constraints (Yang et al., 2000). Furthermore, it has been proposed that most of the proteins appear to be under purifying selection exhibiting dN/dS ratios between 0 and 1 (Li, 1997), and therefore adaptive evolution most likely occurs at a few amino acids (Yang et al., 2000). However, because this ω ratio is an average over all amino acid sites, it tends to underestimate the positive selection signals. So, an alternative approach was used to search for a molecular signature of positive selection acting on sites within the McyE and NadF coding region.

For this analysis, a number of models (statistical distribution) were compared using the likelihood-ratio criterion for significance. The models used were Model M1, M2, M3, M7, M8 and the previously used M0. M0 represents the hypothesis that all amino acid positions exhibit the same ω. Model M1 (neutral model) represents the hypothesis that all amino acid positions are completely constrained (ω = 0) or neutral (ω = 1). Model M2 is a selection model that allows for an additional category of amino acid sites that may be positively selected (ω > 1). Model M3 (discrete model) uses an unconstrained discrete distribution to model heterogeneous ω ratios among sites. Models M7 (β-model) and M8 (β + ω model) are similar, both use the distribution of ω as a β-distribution, but M7 is a special case of M8, that include an additional category of sites that are not part of the β-distribution. Therefore, the model M1 (neutral) was compared with the model M2 (selection), the model M3 (discrete) was compared with M0 (same ω), and because the M7 is a special case of M8, the more general model M8 can be compared with M7. When the more general models indicate the presence of sites with ω > 1, the comparison constitutes an LRT of positive selection (Yang et al., 2000).

The likelihood values obtained from PAML for the three pairs of models with different assumed ω distributions were compared using the likelihood ratio test (LRT). Thus, when two models are nested, twice the log-likelihood difference will be compared with a $\chi 2$ distribution with the degrees of freedom (d.f.) equal to the difference in the number of parameters between the two models (Yang et al., 2000). The likelihood results for all models are represented in the Table 5, as well as the amino acid positions that retrieved a significant value for positive selection identified by the PAML analysis.

Table 5. Likelihood ratio tests of positive selection for the McyE estimated in PAML.

Model	lnL	LRT	Estimated parameters	Positive selected sites
M8: beta & ω > 1	1301. 714	6.596 (p < 0.05, d.f. = 2)	p0= 0.985, p= 0.289, q= 3.490, (p1= 0.014), ω= 1.933	(BEB) **72K**, ω= 2.122, PP= 0.950 **73T**, ω= 1.289, PP= 0.528 (NEB) **72K**, ω= 1.930 PP= 0.998
M7: beta	1305.012		p= 0.195 q= 1.696	Not allowed
M2: positive selection	1307.162	0	p0= 0.930, p1= 0.053, p2= 0.016, ω0= 0.045, ω1=1, ω2= 1	(NEB) **72K**, ω= 1.902 PP= 0.746
M1: Nearly neutral	1307.162		p0= 0.930, p1= 0.069, ω0= 0.045, ω1= 1	Not allowed
M3: discrete	1301.600	49.868 (p < 0.05, d.f. = 4)	p0= 0.749, p1= 0.235, p2= 0.015, ω0= 0.015, ω1= 0.245, ω2= 1.911	(NEB) **72K**, ω= 1.910 PP= 0.999 **73T**, ω= 1.129 PP= 0.530
M0: single ω	1326.534		ω= 0. 076	Not allowed

* Log-likelihood scores (InL) were compared for each pair of models (M1 versus M2, M0 versus M3, and M7 versus M8) using the test statistic LRT, with significance evaluated from χ2 distribution; PP: posterior probability.

For each LRT, the models M8 and M3 that allow for sites to be under positive selection fit the data better than the neutral models (M1 or M7) (Table 5). M8 identified two amino acid sites under positive selection using BEB (Bayes Empirical Bayes) (Yang et al., 2005), 72K (PP ≥ 95%) and 73T, and the NEB (Naive Empirical Bayes) identified only the site 72K (PP ≥ 99%). The sites 72 and 73 correspond to the positions 88 and 89 in the alignment, respectively. M8 highlights that most sites within the data set experience purifying selection (ω < 98.5 %). The selection model M2 show no improvement of likelihood over the neutral model, but the discrete model M3 is accepted with a statistical significance p <0.05, and identified the site 72K as a position under positive selection (by NEB) with PP = 0.999.

A plot with the ω value by amino acids positions obtained from the BEB analysis of M8 is represented in Fig.8. This representation shows that most sites are under strong purifying selection, with just a few sites having some relaxation of the constraints forces, but only one site having significant evidence of being under positive selection. Furthermore, the sliding windows analyses of amino acids sequence identity (Fig.9) confirmed these conclusions. The overall pattern is that highly conserved sites appear to be followed by sites with relaxation of selective constraints and there is one site with less than 40% of amino acids identity at the position previously identified as positively selected. Also, the

comparison of the sliding windows analyses of the substitution frequency (dN/dS) with the percentage of similarity (Fig.9) shows that the conservative sites (very low values of dN/dS and highly percent of similarity) are well distributed across the sequence but are interrupted by less conservative sites (having a slightly higher dN/dS and low values of percent identity).

A few sites revealed peaks of non-synonymous substitutions, one corresponding to the position 88 (in the alignment) that was inferred to be under positive selection using PAML (Fig. 9). The other peak of non-synonymous mutations is localized at the end of the sequence, around the position 124 (in the alignment) that match with the positions around 106 in the Fig.8 and show relaxation of the constraint forces with $0 < \omega < 1$. In general the synonymous substitutions have a little effect over ω because there are not under selective pressure (Yang et al., 2000). But there is a very high peak of synonymous mutations that match with a slight increase of non-synonymous mutations (around the positions 66-68 in the Fig.9) that had an effect on ω (Fig. 8 positions 48-52). This reflects the relaxation of constraint forces around these positions to such extend that is observed in the decrease of percent of identity of the amino acids in the protein.

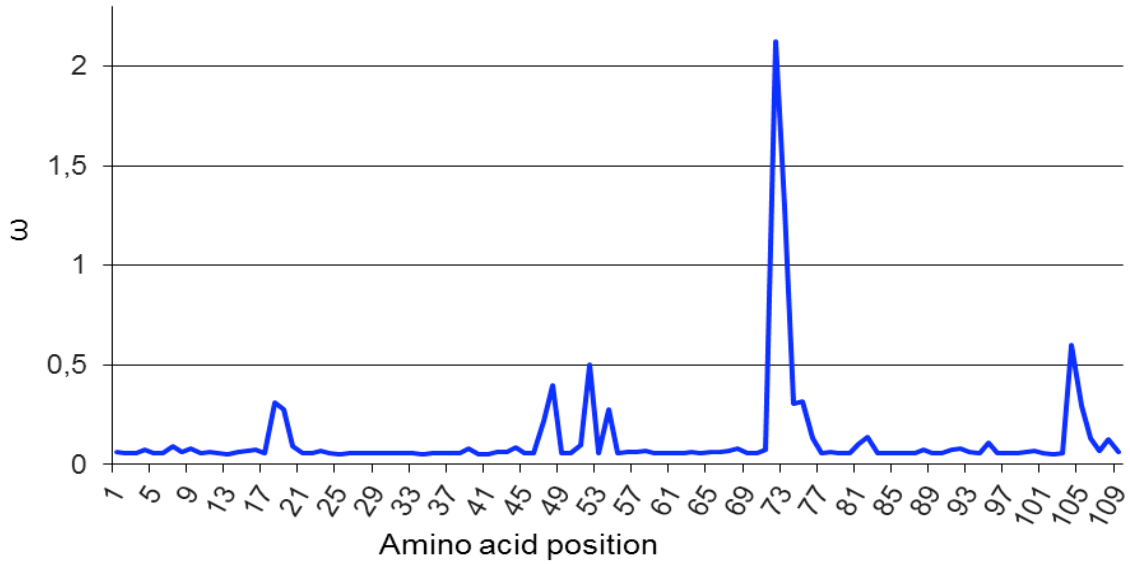

Fig.8. Results of synonymous and non-synonymous ratio (ω) retrieved from the BEB analysis for the McyE gene (labeled by the amino acid position).

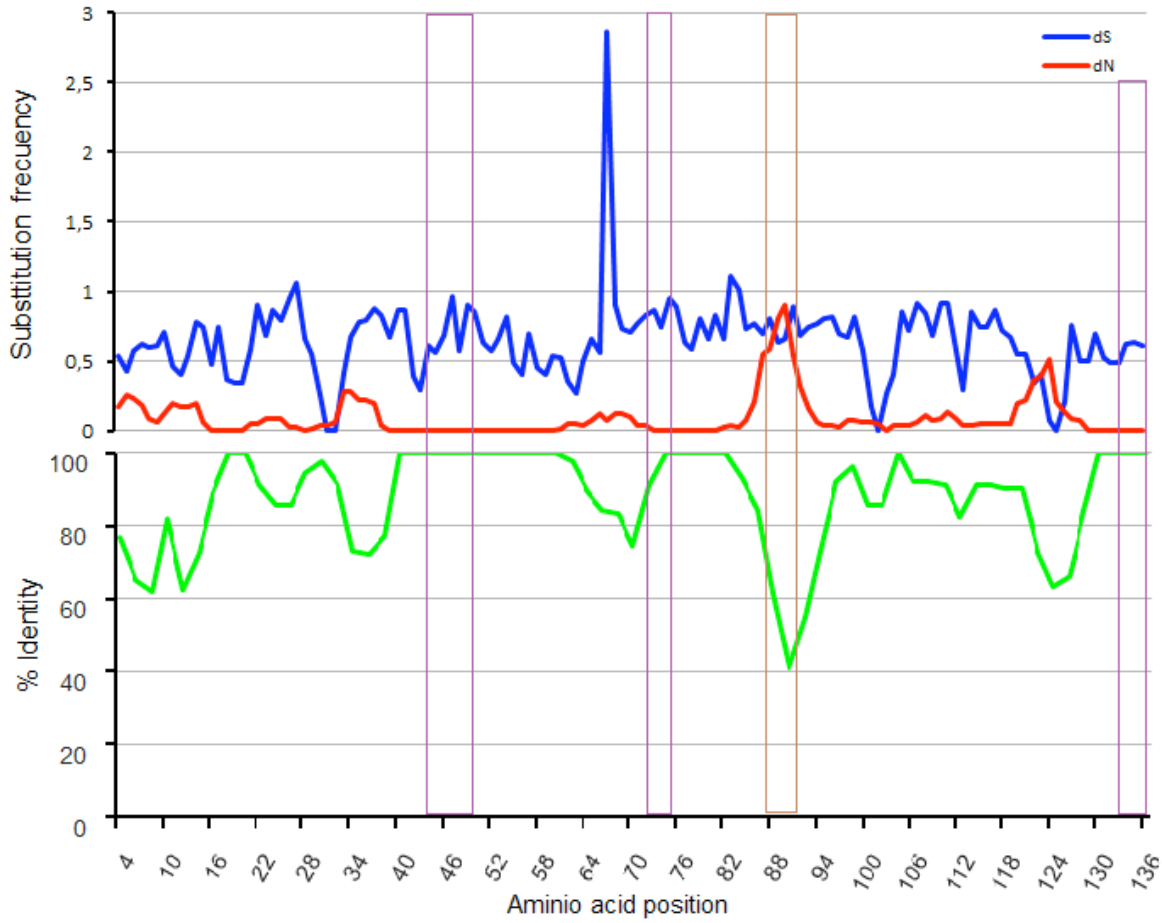

Fig.9. Sliding window analyses of substitution frequency and of amino acid sequence identity by amino acids site in the McyE gene. Colored boxes: Lilac, predicted binding sites of the cofactor; Brown, positive selection site.

In gene sequences the evolutionary nucleotide substitutions that produce amino acid replacements occur less frequently than silent substitutions (Jukes, 1987). These silent nucleotide substitutions in the protein coding-regions are either transitions (purine-purine or pyrimidine-pyrimidine interchanges) or transversions (purine-pyrimidine interchanges) (Jukes, 1987). Under the Jukes nomenclature the transitions and transversions are divided into silent, unconstrained (optional) and constrained (obligate) categories (Jukes, 1982). The unconstrained transitions and transversions occur in silent and no amino acid replacement is produced. In some cases third-position transversions produce amino acid replacement but there are less frequent than transitions.

In the Fig.10 it can be observed that although the transversions are more frequent that the transitions, this is not reflected in an amino acid substitution and therefore there are silent substitutions. But the constraint against the replacement substitutions has become lower at one position that was previously identified as a positively selected site (position 88 and 262-264 in the amino acid and nucleotide alignment respectively), therefore several amino acid replacements are produce. The amino acids that can be found in the position 88 are:

lysine (K), proline (P), valine (V), alanine (A), leucine (L), serine (S), and threonine (T).

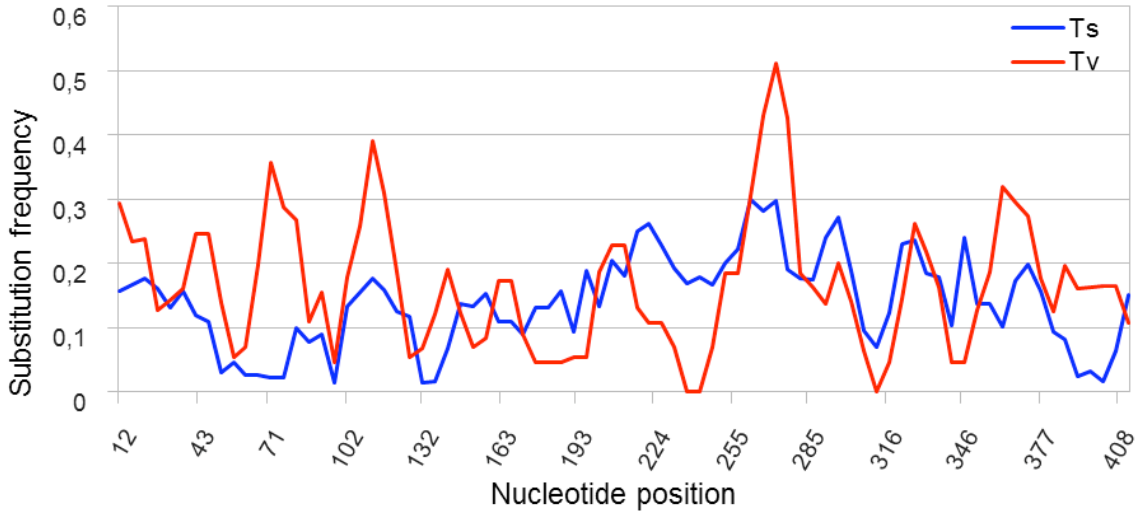

Fig.10. Sliding window analysis of transitions (Ts) and transversions (Tv) estimated in SWAAP for the McyE gene.

3.3.1. - Site by site analysis of natural selection

Statistical inference is never 100% accurate, and there may be some false positives or negatives. So, different approaches to infer the same effect should be made to avoid unreliable inferences (Poon et al., 2009). Therefore, three different codon-based maximum likelihood methods, FEL, REL and SLAC were used for detecting signatures of positive and negative selection from coding sequence alignments in the web server Datamonkey. These programs examine ratios of non-synonymous (dN) and synonymous mutations (dS) to identify signals of positive (dN > dS) and negative (purifiying) selection (dN < dS) operating on individual codons within genes. These methods do not remove ambiguity sites, so the numeration of the alignment is maintained. All three methods successfully identify the 88K as positively selected site, also identified by PAML, but have not identified the site 89T retrieved with low posterior probability in PAML (Table 6).

Table 6. Positively selected site in the McyE gene detected by Datamonkey.

Codon	**SLAC p-value (≤ 0,1)**	**FEL p-value (≤ 0,1)**	**REL Bayes Factor (≥50)**
88K	0.053	0.046	1923040

Since these methods have better results with large data sets (>100 sequences), the authors recommend to run all the three methods and compare their results side-by-side (Poon et al., 2009). Thus, if the methods corroborate each other's findings, it is likely to have more confidence in the inference. Therefore, only the negatively selected sites

detected by all three methods were taken into account (Table 7). The AMT sequences analyzed contain 139 amino acids, for which Datamonkey found one site as positively selected with statistically significance and 31 sites as negatively selected sites. Most sites identified as negatively selected are located in the conserved positions in the sliding windows analysis of substitution frequency and percent of similarity (Fig.9), but some negatively selected sites do not appear to match with these conservative zones.

Table 7. Negatively selected sites in the McyE gene detected by Datamonkey.

Codon	**SLAC p-value (≤ 0,1)**	**FEL p-value (≤ 0,1)**	**REL Bayes Factor (≥50)**
3V	0.037	0.019	81263
8Q	0.001	0.000	1017100
15I	0.001	0.001	1104710
16G	0.014	0.005	227613
24A	0.037	0.013	101543
27T	0.037	0.004	356452
28A	0.037	0.004	368219
35G	0. 088	0.068	1785
39R	0.041	0.007	236024
40V	0.004	0.001	1141600
52A	0.037	0.002	713441
55R	0.014	0.003	857270
59S	0.037	0.003	433714
60R	0.065	0.017	72972
66I	0.021	0.006	179419
73Y	0.059	0.010	74924
74H	0.035	0.015	62874
75G	0.039	0.016	75758
82A	0.012	0.001	2185300
86E	0.024	0.009	83979

96L	0.016	0.000	5309490
97G	0.037	0.009	179836
98T	0.004	0.000	17026500
105D	0.041	0.006	141336
107I	0.048	0.005	127200
116S	0.045	0.010	87285
121A	0.012	0.000	3857020
128A	0.037	0.016	57255
129A	0.037	0.012	75997
134P	0.037	0.001	300881
136Q	0.031	0.002	193183

3.4. - Biochemical properties constraints of the McyE encoded protein

The models implemented by PAML assume that the non-synonymous substitution rate is independent of the amino acid being interchanged, that is, for a given positively selected site, all amino acids are assumed to be advantageous. Therefore, analyses to detect constraints for some biochemical characteristic, such as polarity, volume, and charge, were made to identify positions that experience more non-conservative substitutions, which can be under positive selection.

This analysis performed in ConTest failed to detect the site previously identified as under positive selection by PAML, but was able to identify other sites (seven in total) with a percent of identity around 60%. However, these sites were not supported by sufficient statistical significance (by the Bonferroni correction and the False Discovery Rate). Therefore it cannot be claimed with confidence that such sites are under positive selection. However, this analysis revealed a pattern on the constraints of substitution frequency that may be of interest for functional and/or structural important positions. The biochemical characteristic more constrained was the charge at the sites 5, 6, 9, 13, 123, and 125 while in the position 36 was the volume. Thus, the constraint of the charge at these positions may be important to maintain some functional or structural feature of the protein. To verify this, the sites selected by ConTest were represented in the secondary structure of the AMT domain made in this study. Interestingly the selected sites with the charge as biochemical constrained feature, corresponded to two α-helix in the secondary

structure, which are highly limited by charge and volume of the amino acids for avoid the destabilization of the structure (Fig.11). The position 36 located at the C-terminal end of a α-helix have the volume as biochemical constrained feature, and possibly drive the destabilization of the α-helix by the rupture of the structure, leading to random coil structures. So, it appears that these sites can have a relaxation of the purifying selection but the biochemical characteristics of the amino acids at these sites are maintained in order to avoid the destabilization of the secondary structure of the protein domain.

3.5. - Secondary structure of the AMT domain

The lack of crystal structures for the AMT domains from the PKS/NRPS biosynthetic pathway make difficult to model the secondary structure of the aminotransferase domain of the McyE. It is know that the AMT domain of the McyE is located at the hybrid interface of the PKS-NRPS modules in the McyE (Tillet et al., 2000), a special position in the MC gene cluster. So, a crystalized structure from the well characterized aminotransferase from the *Bacillus subtilis* was used as template for I-Tasser to construct the secondary structure model for the AMT of McyE.

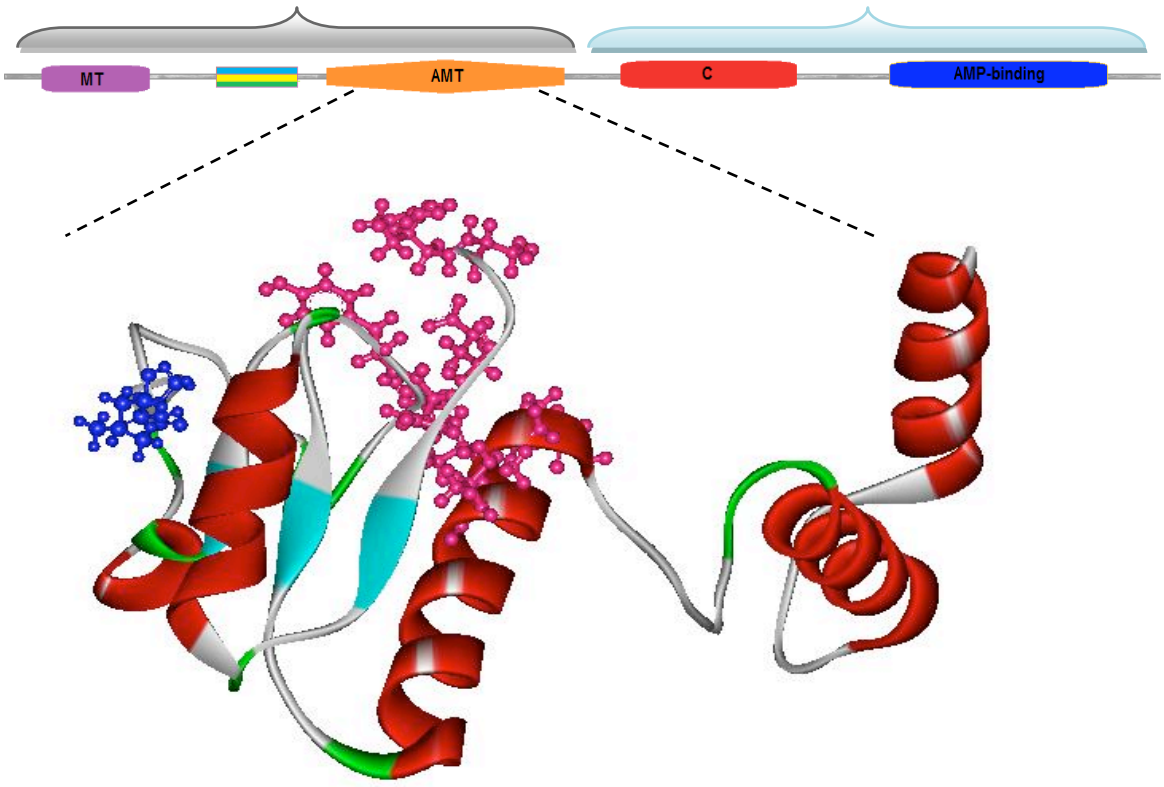

Fig.11. Partial organization of the catalytic domains in McyE and secondary structure of the AMT domain. The PKS and NRPS modules are identified by braces, grey and blue, respectively. In the cartoon of the modules the domains are MT: methyltransferase, AMT: aminotransferase, C: condensation, AMP-binding: AMP binding. In the secondary structure are shown with balls and

sticks both the cofactor binding positions (pink) and the positively selected sites (blue) from the BEB analysis of M8 in PAML.

Curiously, *Bacillus subtilis* produce a secondary metabolite, the micosubtilin, which is produced by a gene cluster named *myc* constituted by four ORFs: *fenF*, *mycA*, *mycB*, *mycC*. The MycA contain an AMT domain that shows significant amino acid similarity (54%) to a pyridoxal 5'-phosphate (PLP)-dependent enzyme family that includes the glutamate semialdehyde aminotransferase (GSA) (Duitman et al., 1999), an aminotransferase class-III. Furthermore, the AMT domain from the mycosubtilin gene cluster (MycA) is also located at the hybrid interface of PKS-NRPS modules. This kind of aminotransferase domain has been only identified in three systems: the iturins (like mycosubtiin), the microcystin and the prodigiosins (Aron et al., 2005).

The cofactor (pyridoxal-5'-phosphate) binding area (Fig. 11) identified by I-Tasser in the secondary structure of the aminotransferase domain of McyE, correspond partially to the cofactor binding area of the aminotransferase model, while the catalytic residue cannot be detected because it is located several positions beyond the end of the sequence positions analyzed in this work.

The residues identified as the predicted binding site for the cofactor in the AMT domain of McyE are the following: 45T, 46G, 47T, 50I, 73Y, 74H, 75G, 133E, 137S, and 138R. These sites are represented in the Fig.11. Also, in the Fig.9 it can be observed the cofactor binding sites, which are located in highly conserved areas of the protein.

Using OCA©, a browser-database for protein structure/function, it was possible to relate the residues that correspond to the cofactor binding site in the aminotransferase template with those in the AMT domain. So, the positions in the AMT domain that have an exact pair in the template and their pairs are: 47T/119T, 73Y/145Y, and 133E/207E. The nature of the interactions between residues and cofactor are well described for the crystal structure of the template. Thus, based in the properties of the crystal structure it can be said that the position 47T/119T form a hydrophilic contact bond with the cofactor. The position 73Y/145Y form a hydrophobic contact bond with the cofactor. The position 133E/207E form a hydrophobic/hydrophilic contact bond with the cofactor that is considered as a destabilizing contact.

On the other hand, in the secondary structure it can be observed that the cofactor binding sites are spatially separated from the putative positively selected site.

4. - Discussion

4.1. - Importance of the AMT and their special features

The McyE and the NdaF are hybrid proteins with a NRPS/PKS system (Moffitt and Neilan, 2004), and their AMT domains have a special localization within the protein and within the microcystin and nodularin gene cluster. The AMT domain is located at the hybrid interface of PKS/NRPS (Tillet et al., 2000) (Fig.12) and is connected to the first module of the peptide synthetase via an extra condensation domain. The AMT domain has homology to a large group of non-integrated semialdehyde aminotrasferases (Tillet et al., 2000) that are widespread in the primary and the secondary metabolism in microorganisms, and includes the glutamate semialdehyde aminotransferase that is involved in the formation of chlorophyll in cyanobacteria (Beale, 1994).

It is noteworthy that the pyridoxal 5'-phosphate (PLP)-dependent domains are rare in thiotemplate assembly lines and have been identified in only three systems: the iturins (like mycosubtiin), the microcystin and the prodigiosins (Aron et al., 2005). So, the AMT domains are located at a special position that makes the connection between the PKS and the NRPS modules but are rare in assembly lines.

The aminotransferase domain is a catalytic unit that catalyzes the transfer of an amine group to the beta position of the growing chain in the metabolic pathway of micosubtilin. This mechanism has been elucidated by Aron et al., (2005) that characterized the AMT domain of this pathway. Aron et al., (2005) found that a reversible amine transfer from the amino acid glutamine (Gln) to enzyme-bound PLP results in the formation of an enzyme-pyridoxamine 5’-phosphate (PMP) complex. Then it occurs the displacement of the amino-acid-derived α-ketoacid by a thiolation domain (ACP)-bound β-ketothioester and the reversible amine transfer generates the corresponding covalently tethered β-aminothioester.

Probably the mechanism whereby the transfer of the amino group on the Adda moiety of the microcystin biosynthetic pathway is made in similar way to the transfer mechanism of the amino group in the micosubtilin pathway. This can be inferred because all aminotransferases use the same cofactor to catalyze the same type of reaction (Mehta et al., 1993), and because there are largely homologies between the AMT domains of micosubtilin and the microcystin.

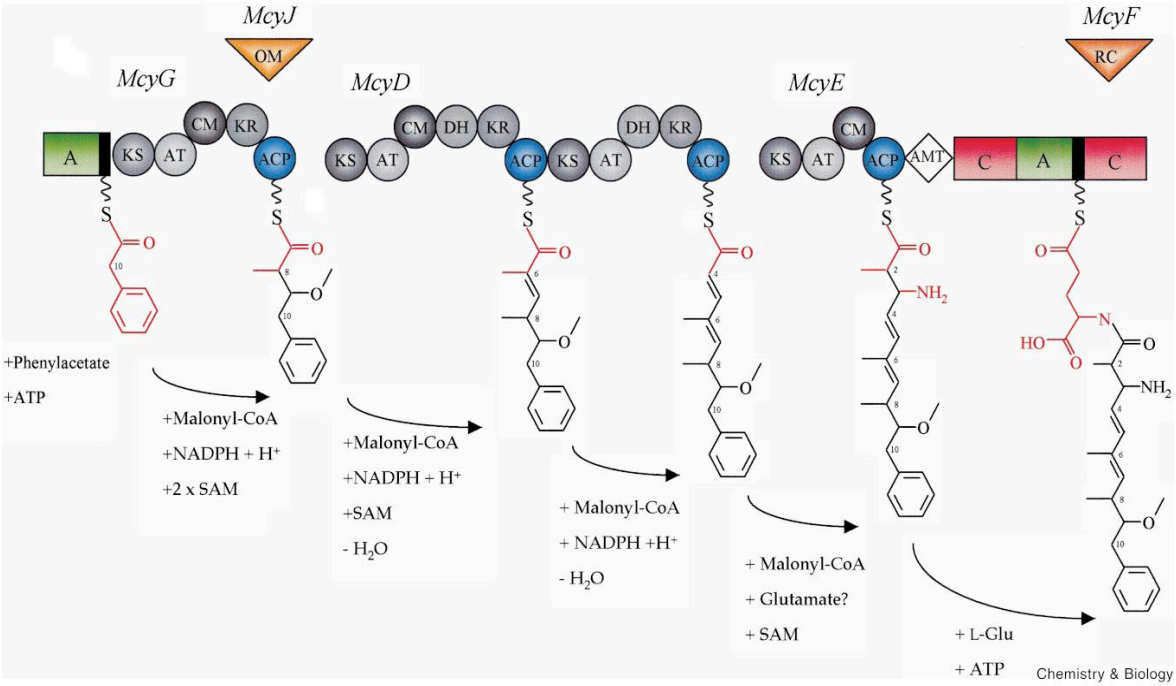

Fig.12. Model for the formation of Adda including the predicted domain structure of McyG, McyD and McyE. PKS domain: circles; NRPS domains: rectangles; Aminotransferase domain: diamond. Tailoring ORFs McyJ and McyF: inverted triangles; KS: L-ketoacyl synthase; AT: acyltransferase; ACP: acyl carrier protein; KR: ketoacyl reductase; DH: dehydratase; CM: C-methyltransferase; OM: O-methyltransferase; A: aminoacyl adenylation; C: condensation; AMT: aminotransferase; RC: racemase. The reaction order shows the transfer and the condensation of Adda to D-glutamate (Tillet at al., 2000).

In the *in silico* study made by Ansari et al. (2008) the arrangement of the AMT domain together with a methyltransferase domain (MT) could only be observed in two examples from the all MT domains analyzed. Such result indicated that the MT domain was adjacent to an aminotransferase (AMT) domain only in hybrid NRPS/PKS system, and in these cases MT-AMT stretch was inserted between a PKS and a NRPS module. Thus, it was inferred that the specific location of the AMT domain might be significant for functional communication with the catalytic modules upstream and downstream (Aron et al., 2005). Furthermore, it can be an interest point of insertion of additional domains in the assembly line that open new opportunities in the rational engineering of the NRPS-PKS systems to produce novel metabolites as has been proposed a decade ago by Tillet et al. (2000).

4.2. - Selective forces acting in the AMT domain

Under the population genetics model of molecular evolution (Kimura, 1983), the “evolutionary function” is defined as a population genetics parameter that contributes to the fitness of the organism, which is related to the function (biochemical, structural or phenotypic) of a gene in complex ways. It involves interactions at different levels of

biological organization like molecular complexes, pathways and cells, but may also include individuals, population and species.

For the gene cluster of secondary metabolites like MC their evolution is likewise connected to the modularity of their constituent subclusters (Fischbach et al., 2008), therefore all the intermediates of the gene cluster must produce a molecule that justifies the cost of their existence. This may be a “functionalistic view” considering the fact that the number of secondary metabolites drastically outnumbers those with an assigned function (Jenke-Kodama and Dittmann, 2009). Otherwise, the intermediate enzymes may have not been necessary for the functioning of the cluster, so the evolutionary origin of this kind of clusters is largely unclear (Fischbach et al., 2008).

As showed previously for other segments of the *mcy* gene cluster (Tanabe et al., 2004; Rantala et al., 2004; Tooming-Klunderud et al., 2008), purifying selection maintains most part of the sequence under selective constraint indicating that mutations that affect the amino acid sequence of the domain are generally deleterious. So, as it is shown in the Fig.8 a strong functional constraint is responsible for the general low dN/dS ratio across the AMT domain, which maintains the long-term stability of the protein. While one site of the AMT domain stands out with significant signals of positive selection despite the relative conservation of the rest of the sequence. Due to the important function of McyE in the assembly of the MC and given that the gene is not involved in the production of MC isoforms, it was expected that the AMT domain would be a highly conserved motif. However, it has been shown that during sequence evolution a particular amino acid residue can be changed from very conserved to highly variable and *vice versa* (Gu, 2003), and that may be an indicative of functional divergence. Tooming-Klunderud et al. (2008) found evidence for positive selection in the adenylation domains of the *mcyB* and the *mcyC* in *Microcystis*, *Planktothrix* and *Anabaena*. These domains are involved directly in the recognition and activation of amino acids in the MC and therefore in the production of MC isoforms. However, they also found that in the first adenylation domain of McyB there are several positively selected amino acid residues that are not associated with substrate selectivity, concluding that some other property is selected for these adenylation domains. They suggest that this could be due to changes in the catalytic efficiency or in the interactions between neighboring domains and modules. Since the AMT domain is not associated to substrate selectivity and is an important link between modules (and therefore for their interactions), then it is possible that some unknown functionality could be under selection for the AMT domain like in the adenylation domains.

The positive selection act by increasing rare variants that improve optimal fitness. However, is has been demonstrated that the lack of microcystin production does not imply a selective disadvantage for the individuals, so, the metabolic cost of the MC synthesis

are not necessarily of relevance for the survival of the individuals. This is shown by the natural coexistence of microcystin-producing and non-producing strains (Christiansen et al., 2008), the demonstration of the inactivation of the MC synthesis did not increase the growth under different conditions (Hesse et al., 2001), and also that so far has not been assigned any function to MC. Thus, the possession of MC genes might be favorable in certain environmental conditions for a specific strain (Nishizawa et al., 2000).

On the other hand, it has been proposed that due to the difference in the growth rate of strains under laboratory conditions, the measure of this parameter cannot be used to relate the fitness of a particular strain to the presence or absence of MC (Christiansen et al., 2008). Due to this kind of experimental inaccuracy it can be said that up to now there are no reliable data on the estimation of fitness cost caused by secondary metabolites (like microcystin) in bacteria or in microbial populations (Jenke-Kodama and Dittmann, 2009).

4.3. - An integrative point of view

Before discussing the evolution of collective genes or the evolution of one gene that belongs to a gene cluster, it is very important to know more about the gene functional role, genetic characteristics and organization. So such studies together whit available genetic and ecologic information may allow to understand better the evolutionary history of the microcystin synthetase cluster. However, as the AMT domain is part of a metabolic pathway it is necessary not only to look at the processes that shape the biosynthetic enzyme, but also look to the metabolic integration into the overall biochemical network, as suggested by Jenke-Kodama and Dittmann (2009). Thus, the collective nature of the genes in the microcystin pathway cannot be excluded of any approach that tries to make conclusions about the evolution of the MC cluster and their hypothetic function under natural conditions.

To hold this proposal, it has been suggested that both the degree of constraint and the rate of adaptive substitution in proteins, and hence the overall rate of substitution, are likely influenced not just by intrinsic properties of particular proteins, but also by their network properties (Rausher et al., 2008). Some of these network properties that can influence the evolution of a protein that belong to a metabolic pathway are resumed by Rausher et al. (2008) as: the position in a metabolic network, level of expression, and the number of interacting protein partners. Also, it has been proposed that the adaptive substitution are expected to be more common in enzymes at major pathway branch points than in downstream enzymes (Rausher et al., 2008), i.e. in the branching pathways the flux allocation is controlled primarily by enzymes at pathway branch points, so it is expected that natural selection primarily targets enzymes with greatest control on flux. If

this is true, may be a similar approach could be applied to the MC biosynthetic pathway. This approach should take into account the vital role of the Adda moiety for the formation of the MC molecule. The intermediated steps that catalyze their inclusion can be also an important point of control and likely susceptible to natural selection. Thus, the evaluation of the genetic characteristics of *mcyG*, *D* and *E* genes that catalyze the inclusion of Adda would be a proper approach to evaluate this hypothesis. Nevertheless, the importance of the other enzymes that catalyze any other step in the MC assembly cannot be a priori excluded.

4.4. - How to proceed?

Understanding the evolutionary origins of the AMT domain is particularly relevant because it is a key domain in the structure of the McyE that facilitate the subcluster fusion. As more gene clusters are sequenced more accurate will be the approaches for reveal the evolutionary history of these particular integrated systems that produce a harmful secondary metabolite microcystin. The direct connection between the gene sequence and the small molecule of microcystin provide an opportunity to study the mechanism by which this systems act. But looking to the current proteins complex, and therefore to the complex product, it is difficult to imagine how this structures could have came through evolution by involving less complex intermediate forms that could not had a function. An approach for systematically address this issue has been proposed by Fischbach et al. (2008). There are two important questions that should be made at the moment of constructing strategies for studing the metabolic pathways of secondary metabolites encoded by gene clusters: (i) if the gene clusters are deconstructed into fragment for analysis, then these fragments should be subclusters, individual genes or sub-portions of genes?, and (ii) how can phylogenies of these fragments be combined coherently to accurately reflect the evolutionary history of a gene cluster? These are important tasks for ultimately try to answer the Darwin's question: how complexity evolved through natural selection?

5. - Conclusions

Taking into account the results of this work and the current knowledge about microcystins, the main conclusions of this work are resumed below.

- The AMT domains of the microcystin gene cluster seem to be primarily under purifying selection, as shown previously for this segment of the *mcy* gene cluster, indicating that mutations that affect the amino acid sequence of this domain are generally deleterious.
- It has been observed a relaxation of the selective constraint in some regions of the AMT domain that have ω values between 0.2 - 0.6. This less conservative regions occurred at intervals interrupting the more conserved regions, which are related to the binding site of the cofactor pyridoxal-5'-phosphate.
- In this study it is provided significant evidence of positive selection acting on one amino acid residue not directly related with the amino acid incorporated into the microcystin molecule. This finding could indicate that some other functionality could be under selection at this site, namely the change in the catalytic efficiency or the interactions between neighboring domains and modules.
- A unique recombination event was found in the date set, suggesting that point mutations are the main cause of genetic variation in the AMT domain of McyE from the genera analyzed in this work.
- The phylogeny of the AMT domains studied in this work is congruent with the results of Rantala et al., (2004), suggesting an ancient origin of the microcystin synthetase genes.
- The results of this work provide insight to understand how the microcystins give arise and how are they maintained through the evolution.
- It is highly recommended to be very careful when trying to draw conclusions from deconstructions of very complex systems that drive the production of secondary metabolites, which should take into account the collective nature of these systems.

6. - References

Abascal, F., Zardoya, R., Posada, D. 2005. ProtTest: Selection of best-fit models of protein evolution. Bioinformatics: 21(9):2104-2105.

Anisimova, M., Bielawski, J. P., Yang, Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol. 18:1585–1592.

Anisimova, M., Bielawski, J. P., Yang, Z. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol Biol Evol. 19:950–958.

Ansari, M. Z., Sharma, J., Gokhale, R. S., Mohanty, D. 2008. *In silico* analysis of methyltransferase domains involved in biosynthesis of secondary metabolites. BMC Bioinformatics. 9:454.

Ansari, M. Z., Yadav, G., Gokhale, R. S., Mohanty, D. 2004. NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Res. 32:W405–W413.

Aron, Z. D., Dorrestein, P. C., Blackhall, J. R., Kelleher, N. L., Walsh, C. T. 2005. Characterization of a new tailoring domain in polyketide biogenesis: the amine transferase domain of MycA in the mycosubtilin gene cluster. J Am Chem Soc. 127(43):14986-14987.

Barker, G. L., Handley, B. A., Vacharapiyasophon, P., Stevens, J. R., Hayes, P. K. 2000. Allele-specific PCR shows that genetic exchange occurs among genetically diverse *Nodularia* (cyanobacteria) filaments in the Baltic Sea. Microbiology. 146:2865-2875.

Beale, S. I. 1994. Biosynthesis of cyanobacterial tetrapyrrole pigments hemes, chlorophylls, and phycobilins. In: Bryant, D. A (Ed). The molecular biology of cyanobacteria (Advances in Photosynthesis). Vol 1. Kluwer, The Netherlands. pp 519-553.

Black, T. A., Wolk, C. P. 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(8): 2282-2292.

Broadhurst, R. W., Nietlispach, D., Wheatcroft, M. P., Leadlay, P. F., Weissman, K. J. 2003. The structure of docking domains in modular polyketide synthases. Chem Biol. 10:723-731.

Bruen, T., Phillipe, H., Bryant, D. 2006. A quick and robust statistical test to detect the presence of recombination. Genetics. 172:2665-2681.

Caboche, S., Pupin, M., Leclère, V., Jacques, P., Kucherov, G. 2009. Structural pattern matching of nonribosomal peptides. BMC Structural Biology. 9:15.

Chessel, D., Dufour, A. B., Thioulouse, J. 2004. The ade4 package-I- One-table methods.

R News. 4:5-10.

Christiansen, G., Fastner, J., Erhard, M., Börner, T., Dittmann, E. 2003. Microcystin biosynthesis in *Planktothrix*: genes, evolution, and manipulation. J Bacteriol. 185(2):564-572.

Christiansen, G., Kurmayer, R., Liu, Q., Börner, T. 2006. Transposons inactivate biosynthesis of the nonribosomal peptide microcystin in naturally occurring *Planktothrix* spp. Appl Environ Microbiol. 72(1):117-123.

Christiansen, G., Molitor, C., Philmus, B., Kurmayer, R. 2008. Nontoxic strains of cyanobacteria are the result of major gene deletion events induced by a transposable element. Mol. Biol. Evol. 25(8):1695-1704.

Czárán, T. L., Hoekstra, R. F., Pagie, L. 2002. Chemical warfare between microbes promotes biodiversity. Proc. Natl. Acad. Sci. 99(2):786-790.

Duitman, E. H., Hamoen, L. W., Rembold, M., Venema, G., Seitz, H., Saenger, W., Bernhard, F., Reinhardt, R., Schmidt, M., Ullrich, C., Stein, T., Leenders, F., Vater, J. 1999. The mycosubtilin synthetase of *Bacillus subtilis* ATCC6633: a multifunctional hybrid between a peptide synthetase, an amino transferase, and a fatty acid synthase. Proc Natl Acad Sci. 96(23):13294-13299.

Dittmann, E., Börner, T. 2005. Genetic contributions to the risk assessment of microcystin in the environment. Toxicol Appl Pharmacol. 203(3):192-200.

Dittmann, E., Neilan, B. A., Erhard, M., von Döhren, H., Börner, T. 1997. Insertional mutagenesis of a peptide synthetase gene that is responsible for hepatotoxin production in the cyanobacterium Microcystis aeruginosa PCC 7806. Mol Microbiol. 26(4):779-87.

Du, L., Shen, B. 2001. Biosynthesis of hybrid peptide-polyketide natural products. Curr Opin Drug Discov Devel. 4(2):215-228.

Du, L., Cheng, Y. Q., Ingenhorst, G., Tang, G. L., Huang, Y., Shen, B. 2003. Hybrid peptide-polyketide natural products: biosynthesis and prospects towards engineering novel molecules. Genet Eng. 25:227-267.

Dutheil J. 2008. Detecting site-specific biochemical constraints through substitution mapping. J Mol Evol. 67(3):257-365.

Fewer, D. P., Köykkä, M., Halinen, K., Jokela, J., Lyra, C., Sivonen K. 2009. Culture-independent evidence for the persistent presence and genetic diversity of microcystin-producing *Anabaena* (Cyanobacteria) in the Gulf of Finland. Environ Microbiol. 11(4):855-866.

Fewer, D. P., Rouhiainen, L., Jokela, J., Wahlsten, M., Laakso, K., Wang, W., Sivonen, K. 2007. Recurrent adenylation domain replacement in the microcystin synthetase gene cluster. BMC Evolutionary Biol. 7:183.

Finlay, B. J. 2002. Global dispersal of free-living microbial eukaryote species. Science. 296(5570):1061-1063.

Fischbach, M. A., Walsh, C. T., Clardy, J. 2008. The evolution of gene collectives: How natural selection drives chemical innovation. Proc Natl Acad Sci. 105(12):4601-4608.

Gevers, W., Kleinkauf, H., Lipmann, F. 1968. The activation of amino acids for biosynthesis of gramicidin S. Proc Natl Acad Sci. 60:269.

Gibbs, M. J., Armstrong, J. S., Gibbs, A. J. 2000. Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics. 16:573-582.

Gouy M., Guindon S., Gascuel, O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 27(2):221-224.

Gu, X. 2003. Functional divergence in protein (family) sequence evolution. Genetica. 118:133-141.

Guindon, S., Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biol. 52(5):696-704.

Hesse, K., Dittmann, E., Börner, T. 2001. Consequences of impaired microcystin production for light-dependent growth and pigmentation of *Microcystis aeruginosa* PCC 7806. FEMS Microbiol Ecol. 37(1):39-43.

Huson, D. H., Bryant, D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23(2):254-267.

Jenke-Kodama, H., Dittmann, E. 2009. Evolution of metabolic diversity: Insights from microbial polyketide synthases. Phytochemistry. 70:1858-1866.

Jenke-Kodama, H., Sandmann, A., Müller, R., Dittmann, E. 2005. Evolutionary implications of bacterial polyketide synthases. Mol Biol Evol. 22(10):2027-2039.

Jones, D. T., Taylor, W. R., Thornton, J. M. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275-282.

Jukes, T. H. 1982. Silent nucleotide substitutions in evolution. Paper presented at the Meeting of the Society for Study of Evolution, Stony Brook NY.

Jukes, T. H. 1987. Transitions, transversions, and the molecular evolutionary clock. J Mol Evol. 26:87-98.

Jungblut, A. D., Neilan, B. A. 2006. Molecular identification and evolution of the cyclic peptide hepatotoxins, microcystin and nodularin, synthetase genes in three orders of cyanobacteria. Arch Microbiol. 185(2):107-114.

Kalaitzis, J. A., Lauro, F. M., Neilan, B. A. 2009. Mining cyanobacterial genomes for genes encoding complex biosynthetic pathways. Nat Prod Rep. 26:1447-1465.

Kaebernick, M., Dittmann, E., Börner, T., Neilan, B. A. 2002. Multiple alternate transcripts direct the biosynthesis of microcystin, a cyanobacterial nonribosomal peptide. Appl Environ Microbiol. 68(2): 449-455.

Kaebernick, M., Neilan, B. A. 2001. Ecological and molecular investigations of cyanotoxin production. FEMS Microbiol Ecol. 35:1-9.

Kaebernick, M., Neilan, B. A., Börner, T., Dittmann, E. 2000. Light and the transcriptional response of the microcystin biosynthesis gene cluster. Appl Environ Microbiol. 66(8):3387-3392.

Kaneko, T., Nakajima, N., Okamoto, S., Suzuki, I., and other 19 co-autors. 2007. Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. DNA research. 14:247-256.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, United Kingdom.

Kosakovsky, P. S. L., Frost, D. W. S. 2005a. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments Bioinformatics. 21(10):2531-2533.

Kosakovsky, P. S. L., Frost, D. W. S. 2005b. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 22(5):1208-1222.

Kosakovsky, P. S. L., Frost, D. W. S., Muse, S. V. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 21(5):676-679.

Kurmayer, R., Gumpenberger, M. 2006. Diversity of microcystin genotypes among populations of the filamentous cyanobacteria *Planktothrix rubescens* and *Planktothrix agardhii*. Mol Ecol. 15:3849-3861.

Kurmayer, R., Dittmann, E., Fastner, J., Chorus, I. 2002. Diversity of microcystin genes within a population of the toxic cyanobacterium *Microcystis* spp. in Lake Wannsee (Berlin, Germany). Microb Ecol. 43:107-118.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., Higgins, D. G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23(21):2947-2948.

Li, W. H. 1997. Molecular evolution. Sinauer Associates: Sunderland., MA. USA.

Li, M. H. T., Ung, P. M. U., Zajkowski, J., Garneau-Tsodikova, S., Sherman, D. H. 2009. Automated genome mining for natural products. BMC Bioinformatics. 10:185.

Lodders, N., Stackebrandt, E., Nübel, U. 2005. Frequent genetic recombination in natural populations of the marine cyanobacterium *Microcoleus chthonoplastes*. Environ Microbiol. 7(3):434-442.

Marahiel, M. A., Stachelhaus, T., Mootz, H. D. 1997. Modular peptide synthetases

involved in nonribosomal peptide synthesis. Chem. Rev. (97):2651-2673.

Martin, D. P., Posada, D., Crandall, K. A., Williamson, C. 2005b. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses. 21:98-102.

Martin, D., Rybicki, E. 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics. 16:562-563.

Martin, D. P., Williamson, C., Posada, D. 2005a. RDP2: recombination detection and analysis from sequence alignments. Bioinformatics. 21:260-262.

Mbedi, S., Welker, M., Fastner, J., Wiedner, C. 2005. Variability of the microcystin synthetase gene cluster in the genus *Planktothrix* (Oscillatoriales, Cyanobacteria). FEMS Microbiol Lett. 245(2):299-306.

Mehta, P. K., Hale, T. I., Christen, P. 1993. Aminotransferases: demonstration of homology and division into evolutionary subgroups. Eur J Biochem. 214(2):549-561.

Mikalsen, B., Boison, G., Skulberg, O. M., Fastner, J., Davies, W., Gabrielsen, T. M., Rudi, K., Jakobsen, K. S. 2003. Natural variation in the microcystin synthetase operon *mcyABC* and impact on microcystin production in *Microcystis* strains. J bacteriol. 185(9):2774-2785.

Minowa, Y., Araki, M., Kanehisa, M. 2007. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. J. Mol. Biol. 368:1500-1517.

Moffitt, M. C., Neilan, B. A. 2004. Characterization of the nodularin synthetase gene cluster and proposed theory of the evolution of cyanobacterial hepatotoxins. Appl Environ Microbiol. (11):6353-6362.

Nei, M., Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3(5):418-426.

Nishizawa, T., Asayama, M., Fujii, K., Harada, K., Shirai, M. 1999. Genetic analysis of the peptide synthetase genes for a cyclic heptapeptide microcystin in *Microcystis* spp. J. Biochem. 126:520-529.

Nishizawa, T., Nishizawa, A., Asayama, M., Harada, K., Shirai, M. 2007. Diversity within the microcystin biosynthetic gene clusters among the genus *Microcystis*. Microbes Environ. 22:380-390.

Nishizawa,T., Ueda, A., Asayama, M., Fujii, K., Harada, Ochi, K., Shirai, M. 2000. Polyketide synthase gene coupled to the peptide synthetase module involved in the biosynthesis of the cyclic heptapeptide microcystin. J Biochem. 127:779-789.

Otsuka, S., Suda, S., Li, R., Watanabe, M., Oyaizu, H., Matsumoto, S., Watanabe, M. M. 1999. Phylogenetic relationships between toxic and non-toxic strains of the genus *Microcystis* based on 16S to 23S internal transcribed spacer sequence. FEMS

Microbiol Lett. 172:15-21.

Padidam, M., Sawyer, S., Fauquet, C. M. 1999. Possible emergence of new geminiviruses by frequent recombination. Virology. 265:218-225.

Pearson, L. A., Hisbergues, M., Börner, T., Dittmann, E., Neilan, B. A. 2004. Inactivation of an ABC transporter gene, *mcyH*, results in loss of microcystin production in the cyanobacterium *Microcystis aeruginosa* PCC 7806. Appl Environ Microbiol. 70(11): 6370–6378.

Poon, A. F., Frost, S. D., Pond, S. L. 2009. Detecting signatures of selection from DNA sequences using Datamonkey. Methods Mol Biol. 537:163-183.

Posada, D., Crandall, K. A. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics. 14(9):817-818.

Posada, D., Crandall, K. A. 2001. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc Natl Acad Sci. 98:13757-13762.

Pride, D.T., 2000. SWAAP-a tool for analysing substitutions and similarity in multiple alignments. (http://asiago.stanford.edu/SWAAP/SwaapPage.htm).

Rantala, A., Fewer, D. P., Hisbergues, M., Rouhiainen, L., Vaitomaa, J., Börner, T., Sivonen, K. 2004. Phylogenetic evidence for the early evolution of microcystin synthesis. Proc Natl Acad Sci. 101(2):568-573.

Rausher, M .D., Lu, Y., Meyer, K. 2008. Variation in constraint versus positive selection as an explanation for evolutionary rate variation among anthocyanin genes. J Mol Evol. 67(2):137-44.

Richter, C. D., Stanmore, D. A., Miguel, R. N., Moncrieffe, M. C., Tran, L., Brewerton, S., Meersman, F., Broadhurst, R. W., Weissman, K. J. 2007. Autonomous folding of interdomain regions of a modular polyketide synthase. FEBS J. 274(9):2196-209.

Roberts, M. S., Cohan, F. M. 1995. Recombination and migration rates in natural populations of *Bacillus subtilis* and *Bacillus mojavensis*. Evolution. 49(6):1081-1094.

Ronquist, F., Huelsenbeck, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19(12):1572-1574.

Rouhiainen, L., Vakkilainen, T., Siemer, B. L., Buikema, W., Haselkorn, R., Sivonen, K. 2004. Genes coding for hepatotoxic heptapeptides (microcystins) in the cyanobacterium *Anabaena* strain 90. Appl Environ Microbiol. 70:686-692.

Rounge, T, B., Rohrlack, T., Nederbragt, A. J., Kristensen, T., Jakobsen, K. S. 2009. A genome-wide analysis of nonribosomal peptide synthetase gene clusters and their peptides in a *Planktothrix rubescens* strain. BMC Genomics.10:396.

Roy, A., Kucukural, A., Zhang, Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protocols. 5:725-738.

Schwarzer, D., Finking, R., Marahiel, M. A. 2003. Nonribosomal peptides: from genes to products. Nat Prod Rep. 20:275-287.

Schwarzer, D., Marahiel, M. A. 2001. Multimodular biocatalysts for natural product assembly. Naturwissenschaften. 88:93-101.

Sieber, S. A., Linne, U., Hillson, N. J., Roche, E., Walsh, C. T., Marahiel M. A. 2002. Evidence for a monomeric structure of nonribosomal peptide synthetases. Chem Biol. 9(9):997-1008.

Sieber, S. A., Marahiel, M. A. 2005. Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics. Chem. Rev. 105:715-738.

Smith, J. M. 1992. Analyzing the mosaic structure of genes. J Mol Evol. 34:126-129.

Smith, S. 2002. Modular NRPSs are monomeric. Chem Biol. 9(9):955-959.

Tamura, K., Dudley, J., Nei, M., Kumar, S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 24:1596-1599.

Tanabe, Y., Kasai, F., Watanabe, M. M. 2007. Multilocus sequence typing (MLST) reveals high genetic diversity and clonal population structure of the toxic cyanobacterium *Microcystis aeruginosa*. Microbiol. 153:3695-3703.

Tanabe, Y., Kaya, K., Watanabe, M. M. 2004. Evidence for recombination in the microcystin synthetase (*mcy*) genes of toxic cyanobacteria *Microcystis* spp. J Mol Evol. 58:633-641.

Tanabe, Y., Sano, T., Kasai, F., Watanabe, M. M. 2009. Recombination, cryptic clades and neutral molecular divergence of the microcystin synthetase (*mcy*) genes of toxic cyanobacterium *Microcystis aeruginosa*. BMC Evolutionary Biology. 9:115.

Tillett, D., Dittmann, E., Erhard, M., von Döhren, H., Börner, T., Neilan, B. A. 2000. Structural organization of microcystin biosynthesis in *Microcystis aeruginosa* PCC7806: an integrated peptide-polyketide synthetase system. Chemistry & Biology. 7:753-764.

Tillett, D., Parker, D. L., Neilan, B. A. 2001. Detection of toxigenicity by a probe for the microcystin synthetase a gene (*mcyA*) of the cyanobacterial genus *Microcystis*: comparison of toxicities with 16S rRNA and phycocyanin operon (phycocyanin intergenic spacer) phylogenies. Appl. Environ. Microbiol. 67(6):2810-2818.

Tooming-Klunderud, A., Fewer, D. P., Rohrlack, T., Jokela, J., Rouhiainen, L., Sivonen, K., Kristensen, T., Jakobsen, K. S. 2008. Evidence for positive selection acting on microcystin synthetase adenylation domains in three cyanobacterial genera. BMC Evol Biol. 8:256.

Vasconcelos, V., Martins, A., Vale, M., Antunes, A., Azevedo, J., Welker, M., Lopez, O., Montejano, G. 2010. First report on the occurrence of microcystins in planktonic cyanobacteria from Central Mexico. Toxicon. 56(3):425-431.

Weber, T., Rausch, C., Lopez, P., Hoof, I., Gaykova, V., Huson, D. H., Wohlleben, W. 2009. CLUSEAN: A computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J Biotech. 140:13-17.

Welker, M., von Döhren, H. 2006. Cyanobacterial peptides - Nature's own combinatorial biosynthesis. FEMS Microbiol Rev. 30:530-563.

Wenzel, C. S., Müller, R. 2005. Formation of novel secondary metabolites by bacterial multimodular assembly lines: deviations from textbook biosynthetic logic. Curr Opin Chem Biol. 9:447-458.

Wood, S. A., Heath, M. W., Holland, P. T., Munday, R., McGregor, G. B., Ryan, K. G. 2010. Identification of a benthic microcystin-producing filamentous cyanobacterium (Oscillatoriales) associated with a dog poisoning in New Zealand. Toxicon. 55(4):897-903.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13(5):555-556.

Yang, Z., Nielsen, R., Goldman, N., Pedersen, A. M. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155(1):431-449.

Yang, Z., Wong, W. S., Nielsen, R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107-1118.

Zhang, Y. 2009. I-TASSER: Fully automated protein structure prediction in CASP8. Proteins. S9:100-113.

Zhang, Y. 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 9:40.

Zhao, J., Yang, N., Zeng, R. 2008. Phylogenetic analysis of type I polyketide synthase and nonribosomal peptide synthetase genes in Antarctic sediment. Extremophiles. 12:97-105.