**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Genômica comparativa de *Microcystis aeruginosa* (Cyanobacteria:
Chroococcales), com ênfase em genes envolvidos com síntese de produtos
naturais**

Bruno Weiss

Dissertação apresentada para obtenção do título
de Mestre em Ciências. Área de concentração:
Microbiologia Agrícola

**Piracicaba
2017**

Bruno Weiss
Bacharel em Ciências Biológicas

**Genômica comparativa de *Microcystis aeruginosa* (Cyanobacteria: Chroococcales), com
ênfase em genes envolvidos com síntese de produtos naturais**
versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientadora:
Profa. Dra. **MARLI DE FÁTIMA FIORE**

Dissertação apresentada para obtenção do título
de Mestre em Ciências. Área de concentração:
Microbiologia Agrícola

Piracicaba
2017

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Weiss, Bruno

Genômica comparativa de *Microcystis aeruginosa* (Cyanobacteria: Chroococcales), com ênfase em genes envolvidos com síntese de produtos naturais / Bruno Weiss. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2017.

83 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura "Luiz de Queiroz"

1. Genômica comparativa 2. Cianobactérias 3. Metabolismo secundário 4. Interferência de métodos genômicos I. Título

Aos que ainda mudarão conosco -

Carlos, Isabela, Lara, Zé -

ofereço.

Aos que não mais o farão -

Leonardo, Yves, Zuleika, Vilma, Ursa, Luís -

dedico.

AGRADECIMENTOS

Agradeço à minha mãe e à minha irmã Nara, pela demonstração de que a constância e permanência são a arte que enobrece a consciência.

Ao meu pai, pela demonstração de que o abstrato e o teórico podem ser substanciais e presentes, para a mente e para a percepção apropriadas.

À Profa. Dra. Marli de Fátima Fiore pelos ensinamentos, pela orientação, experiência enriquecedora, e oportunidade de retorno à academia.

Atenciosamente ao amigo Danilo Oliveira de Alvarenga, pelas sugestões, pelo incentivo, companheirismo e exemplo. Desejo sorte em sua quixotesca busca pela abolição dos *Mouses* e apetrechos afins. (Atenção: nenhum animal foi ferido durante o desenvolvimento destas discussões)

Aos colegas do Laboratório de Ecologia Molecular de Cianobactérias - Rafael, Stella, Bruno, Ana Paula, Paula, Janaína, Karina - pelo companheirismo, amizade e paciência.

Ao Programa de Pós-Graduação em Microbiologia Agrícola pelo apoio e organização.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, pela bolsa de estudos.

Desejo a todos, humilde e sinceramente, que a Natureza nos seja evidente.

“Uma obra nunca está terminada.

Ela somente está abandonada.”

Leonardo da Vinci

SUMÁRIO

RESUMO.....	8
ABSTRACT.....	9
1. INTRODUÇÃO.....	11
2. REVISÃO BIBLIOGRÁFICA.....	13
2.1. O Filo Cyanobacteria.....	13
2.2. Metabolismo secundário e vias de síntese de moléculas bioativas em cianobactérias	15
2.3. Via da Sintetase de Peptídeos Não Ribossomais (NRPS).....	17
2.4. Genômica de cianobactérias.....	19
3. HIPÓTESES.....	23
4. OBJETIVOS.....	25
5. MATERIAIS E MÉTODOS.....	27
5.1. Origem das linhagens.....	27
5.2. Análises genômicas e estatísticas, localização e anotação dos agrupamentos relacionados com o metabolismo secundário.....	27
6. RESULTADOS E DISCUSSÃO.....	33
6.1. Agrupamentos gênicos relacionados a produtos naturais encontrados na espécie <i>Microcystis aeruginosa</i>	33
6.2. Comparações do metabolismo secundário entre as linhagens.....	34
6.2.1. Variação dos agrupamentos relacionados a metabólitos secundários mais frequentemente encontrados nos genomas.....	34
6.2.2. Aeruginosina.....	35
6.2.3. Microcistina.....	37
6.2.4. Micropeptina.....	39
6.3. Comparações genômicas globais.....	43
6.3.1. Colinearidade entre as linhagens da espécie <i>Microcystis aeruginosa</i>	44
6.3.2. Homologia de sequência genômica entre as linhagens da espécie <i>M. aeruginosa</i>	47
6.3.3. Estimativa quantitativa do pangenoma e do genoma cerne da espécie <i>M. aeruginosa</i>	49
6.4. Descrição estatística, representatividade dos genomas, e inferências.....	51

6.4.1. Descrição estatística amostral, estimativas da população e correlações.....	54
6.4.2. Correlações das estatísticas descritivas amostrais.....	59
6.4.3. Aumento do genoma mediado por transposases.....	62
6.4.4. Interferência dos métodos genômicos nos estudos comparativos.....	63
7. CONCLUSÕES.....	69
REFERÊNCIAS.....	73

RESUMO

Genômica comparativa de *Microcystis aeruginosa* (Cyanobacteria: Chroococales), com ênfase em genes envolvidos na síntese de produtos naturais#

A ampla diversidade metabólica das cianobactérias é associada não somente a sua importância nos ciclos biogeoquímicos, mas também a sua distribuição global. Tal característica também é responsável pela capacidade destes organismos em produzir uma ampla variedade de substâncias de estruturas incomuns e atividades de interesse para o homem. *Microcystis* é um gênero cianobacteriano reconhecido como produtor de mais de duas centenas de produtos naturais, incluindo cianotoxinas. *Microcystis aeruginosa* é uma espécie frequentemente encontrada em florações, portanto causando preocupações sobre sua influência ecológica, especialmente em corpos d'água doce utilizados para consumo humano. Desta forma, o objetivo deste trabalho foi o levantamento da diversidade e quantidade de metabólitos secundários que podem ser produzidos pela espécie, através de análises genômicas, além de variáveis que podem potencialmente interferir nas análises computacionais, procurando-se por padrões na espécie, e comparando-se 18 linhagens de todos os continentes. Foi encontrado o total de 235 agrupamentos relacionados ao metabolismo secundário, categorizados em 12 classes segundo as estruturas de seus produtos, nas 18 linhagens, evidenciando a riqueza de agrupamentos relacionados ao metabolismo secundário encontrados nesta espécie. Destes agrupamentos, os mais abundantes pertencem às categorias dos Terpenos, Híbridos, Bacteriocinas e NRPS. Entre as NRPS, nenhuma foi comum a todas as linhagens. Ainda, a quantidade de agrupamentos variou entre 6 e 21, e a quantidade de categorias de produtos variou entre 4 e 10, mostrando uma distribuição heterogênea de agrupamentos e tipos de metabólitos preditos. Esta distribuição heterogênea foi detalhada para melhor compreensão deste padrão encontrado na espécie. Dos agrupamentos de NRPS, os três mais frequentes foram selecionados para uma análise pormenorizada de sua estrutura e sequência: aeruginosina (15 linhagens), microcistina (11 linhagens), e micropeptina (15 linhagens). O agrupamento de micropeptina encontrado nas linhagens SPC777, TAIHU98 e PCC 9806 se mostrou amplamente dissimilar com relação à referência utilizada, potencialmente indicando um erro de identificação causado pela plataforma antiSMASH utilizada para a localização dos agrupamentos. Análises de colinearidade genômica mostram uma baixíssima sentença entre os genomas das linhagens em análise, sugerindo frequentes eventos de reorganização genômica. Ainda, análises de pangenoma mostram um cenário em que mais genomas desta espécie são necessários para a estimativa da quantidade total de genes diferentes que a espécie pode possuir, o que é interessante para futuros estudos de procura de metabólitos secundários. Análises do genoma cerne apontam para uma estimativa segura de 1.944 genes comuns a todos os genomas desta espécie, o que corresponde entre 35% e 50% dos genes em cada linhagem. Análises estatísticas apontam para diferentes graus de interferência não linear da quantidade de sequências contíguas na observação de diferentes padrões de outras características genômicas, sugerindo precaução nas expectativas com relação ao metabolismo secundário em caso de linhagens em que a montagem gênica ultrapasse o limite superior aproximado de 100 sequências contíguas.

Palavras-chave: Genômica comparativa; Cianobactérias; Metabolismo secundário; Interferência de métodos genômicos

ABSTRACT

Comparative genomics of *Microcystis aeruginosa* (Cyanobacteria: Chroococales), with emphasis on genes related to natural product synthesis#

The wide metabolic diversity of cyanobacteria is associated not only with their importance in biogeochemical cycles, but also with their global distribution. Such a feature is also responsible for the ability of these organisms to produce a wide variety of substances with unusual structures and activities of interest to man. *Microcystis* is a cyanobacterial genus recognized as a producer of more than two hundred natural products, including cyanotoxins. *Microcystis aeruginosa* is a species frequently found in cyanobacterial blooms, thus causing concerns about its ecological influence, especially in freshwater bodies used for human consumption. In this way, the objective of this work was the survey of the diversity and quantity of secondary metabolites that can be produced by the species, through genomic analyzes, besides variables that can potentially interfere in the computational analyzes, searching for patterns in the species, and comparing 18 strains from all the continents. A total of 235 clusters, categorized in 12 classes according to the structure of their products, were found in the 18 strains, evidencing the richness of clusters related to the secondary metabolism found in this species. Of these clusters, the most abundant belong to the categories of Terpenes, Hybrids, Bacteriocins and NRPS. Among NRPS, none were common to all strains. Also, the number of groups ranged from 6 to 21, and the number of product categories ranged from 4 to 10, showing a heterogeneous distribution of predicted groupings and types of metabolites. Such a heterogeneous distribution was detailed for a better understanding of this pattern found in the species. Of the NRPS clusters, the three most frequent were selected for a detailed analysis of their structure and sequence: aeruginosin (15 strains), microcystin (11 strains), and micropeptin (15 strains). The micropeptide cluster found in the SPC777, TAIHU98 and PCC 9806 strains was widely dissimilar to the reference, so potentially indicating an identification error caused by the antiSMASH platform used to locate the clusters. Genomic collinearity analyzes showed a very low synteny among the genomes of the strains under analysis, suggesting frequent events of genomic reorganization. Also, pangenome analyzes show a scenario in which more genomes of this species are needed for the estimation of the total amount of different genes the species may possess, which is interesting for future studies concerning secondary metabolites. Coregenome analyzes point to a reliable estimate of 1,944 genes common to all genomes of this species, which corresponds to 30% up to 50% of the genes in each strain. Statistical analyzes point to different degrees of non-linear interference of the number of contiguous sequences on the observation of different patterns of other genomic characteristics, suggesting necessary caution about expectations regarding the secondary metabolism in case of strains in which the gene assembly exceeds the approximate upper limit of 100 contiguous sequences.

Keywords: Comparative genomics; Cyanobacteria; Secondary metabolism; Interference of genomic methods

1. INTRODUÇÃO#

As cianobactérias são bactérias gram-negativas metabolicamente diversas, capazes de realizar fotossíntese oxigênica. A distribuição deste filo abrange desde ambientes aquáticos, terrestres e subaéreos, em uma ampla gama de regiões tropicais, temperadas e polares - em sistemas mesófilos, oligotróficos e extremos. Também podem associar-se em relações simbióticas com outros procariotos, plantas, fungos e metazoários (WHITTON; POTTS, 2000; SANTOS; SANT'ANNA, 2010; CASTENHOLTZ, 2001; HARDOIM et al., 2009; KAPLAN et al., 2012). Algumas espécies cianobacterianas são capazes de realizar a fixação de nitrogênio (FIORE; HONDA, 2008). Desta maneira, a sua contribuição relevante para os processos biogeoquímicos da biosfera é, possivelmente, resultado desta combinação de processos metabólicos.

A mesma diversidade metabólica permite também a ocorrência de dominância cianobacteriana em ambientes aquáticos. Tais alterações podem ser resultados de causas naturais ou de causas antrópicas, e são frequentes em ambientes como corpos d'água doce próximos de habitações humanas (DIAS et al., 2004). Os fatores que levam a estas alterações ainda não são consistentemente conhecidos, porém evidências associam florações cianobacterianas a processos de eutrofização por nitrogênio e fósforo (XU et al., 2010). Há também casos em que esta dominância por comunidades cianobacterianas, em certos ambientes, ocorra de forma constante, espontânea e natural, como por exemplo em regiões oceânicas.

As florações são o crescimento demasiadamente rápido de espécies e comunidades cianobacterianas em um corpo d'água. Consequências deste crescimento são o incremento da turbidez da água, hipoxia ou mesmo anoxia do ambiente, e a liberação de moléculas bioativas produzidas pelas cianobactérias (HALLEGRAEFF, 1995). A natureza e atividade de moléculas encontradas numa floração dependem dos grupos cianobacterianos dominantes em cada evento de floração, e da capacidade de cada grupo em produzir tais metabólitos secundários (DITTMAN et al., 2001). Com frequência, as moléculas bioativas liberadas apresentam atividade tóxicas – como saxitoxinas, nodularinas, microviridinas, microcistinas, entre outras. Estas toxinas interferem de diferentes maneiras sobre organismos que entram em contato com a água daquele ambiente, tais como embriófitas, metazoários, eucariontes unicelulares, e bactérias, incluindo mesmo outras cianobactérias (KAEBERNICK; NEILAN, 2001; JONES et al., 2009).

As cianotoxinas são um grupo diverso de moléculas bioativas, possuindo ampla variedade de atividades químicas e bioquímicas. Podem ser substâncias peptídicas cíclicas ou

acíclicas, alcaloides, lipopolissacarídeos, entre outras categorias. Também podem ser peptídeos não ribossomais (NRP), produtos da ação de enzimas megassintetases, comumente enzimas como peptídeo sintetases não ribossômicas (NRPS) e policetídeo sintases (PKS), além de híbridos entre ambas (NRPS/PKS) (DITTMAN et al., 2001; BÖRNER; DITTMAN, 2005).

Linhagens de *Microcystis aeruginosa* são frequentemente produtoras de toxinas, em especial de variedades de microcistinas (FUNARI; TESTAI, 2008; PEARSON et al., 2010). Outros metabólitos secundários também podem ser produzidos, como aeruginosinas e micropeptinas (NAMIKOSHI; RINEHART, 1996). Considerando tal diversidade de substâncias, estudos sobre a produção de moléculas bioativas, especialmente neste grupo, são necessários. A variedade de metabólitos secundários é um aspecto insuficientemente estudado quando consideramos a ampla aplicação e potencial para uso como fonte de moléculas de interesse médico e biotecnológico (MICALLEF et al., 2014). Este fato se dá notoriamente daquelas moléculas bioativas resultantes do sistema NRPS e PKS, devido a sua complexidade e reatividade, além de seu caráter modular e, em alguns casos, potencial preditivo (FELNAGLE et al., 2008; SCHWARZER; FINKING; MARAHIEL, 2002).

A comparação genômica da organização dos genes em agrupamentos de NRPSs e PKSs mostra que a mudança do ordenamento relativo destes genes dentro do agrupamento, assim como translocações, inserções e deleções, podem resultar na produção de moléculas diferentes, capazes tanto de atividades similares das originais, quanto de atividades substancialmente diferentes, dependendo de cada caso (ROUHIINEN et al., 2004), ou ainda a cessação da produção de tais moléculas.

Portanto, fica evidente que, para compreender o sistema e a evolução da produção dos metabólitos secundários, é necessário o entendimento dos padrões e a inferência dos processos que agem internamente na organização dos genes nos agrupamentos responsáveis pela síntese das moléculas bioativas. Da mesma maneira, o entendimento dos processos que levam à modificação da organização do genoma e de seus componentes têm importância no entendimento dos processos locais. Devemos considerar também que genomas procariontes mostram uma grande flexibilidade em sua organização genômica. Estes podem conter poucas regiões cromossômicas conservadas entre linhagens e espécies, enquanto o ordenamento interno dos agrupamentos é, em sua maioria, substancialmente mais conservado em níveis taxonômicos inferiores, tais como espécie (KOONIN, 2009; DARLING, MIKLÓS & RAGAN, 2008).

2. REVISÃO BIBLIOGRÁFICA#

2.1. O Filo Cyanobacteria

As cianobactérias são um grupo monofilético muito diverso de bactérias gram-negativas, fotossintetizantes oxigênicas, que abarca 150 gêneros e cerca de 2000 espécies (APELDOORN et al., 2007). Sua diversidade é considerada parcialmente resultado de sua origem relativamente antiga, a qual remonta a um período de aproximadamente 2,7 Ga – evidenciado por registros geológicos (BROCKS et al., 1999). Contudo, as estimativas de origem do grupo podem variar bastante, algumas mais otimistas alcançando fascinantes 3,5 Ga (SCHOPF, 1996).

Dois tipos de registro geológico são especialmente representativos para a datação do surgimento do grupo, e apontam para um surgimento estimado mais próximo de 2,7 Ga: i) registros fósseis encontrados em estromatólitos australianos, e ii) a detecção de substâncias quimiofósseis 2- α -metilhopanos. Essas substâncias são encontradas em grande parte das cianobactérias atuais, e são responsáveis por conferir rigidez à membrana. A utilidade destas como biomarcadores deve-se a sua recalcitrância, que permite a detecção da molécula mesmo após longos períodos geológicos (BUICK, 1992; SUMMONS et al., 1999; TOMITANI et al., 2006).

Sobretudo, estudos filogenéticos do gene que codifica a subunidade 16S do RNA ribossômico mostram uma origem relativamente tardia do filo Cyanobacteria, quando comparada a alguns outros filós. A divergência do grupo em relação às outras bactérias pode ter ocorrido após a formação de táxons como Thermodesulfobacterales, Chloroflexales, Thermotogales, entre outros grupos de divergência isocrônica (VARGAS et al., 1998). Apesar disso, a definição de grupos internos ao filo Cyanobacteria, valendo-se da mesma técnica de sequenciamento do gene 16S RNAr e comparando com estudos mais antigos de morfologia e bioquímica, apontam para resultados que ainda necessitam de maiores discussões. Ainda, no panorama atual, com o uso de métodos genômicos (o uso de multigenes, filogenômica, e outros), apontam para a necessidade de revisões e atualizações em vários táxons bacterianos, possivelmente abrangendo o filo cianobacteriano.

A longa história evolutiva do grupo pode ser responsável por sua diversidade morfológica, fisiológica e ecológica. Cianobactérias podem ser unicelulares cocoides ou bacilos, ou ainda multicelulares coloniais, filamentosas ou filamentosas ramificadas. Algumas espécies apresentam colônias e células especializadas, como acinetos e heterócitos. Os acinetos são células envolvidas na resistência a condições adversas. Os heterócitos são células diferenciadas, incapazes de realizar fotossíntese oxigênica, de maneira a permitir a atividade

do metabolismo fixador de nitrogênio combinado (CHORUS; BARTRAM, 1999). Essa capacidade de fixação do nitrogênio atmosférico é resultado da presença da enzima nitrogenase, sensível ao oxigênio, e redutora do nitrogênio diatômico à amônia (STEWART, 1973), o que confere relevância a estas espécies tanto no ciclo biogeoquímico do carbono, quanto no ciclo do nitrogênio.

As cianobactérias apresentam importância ecológica e econômica. Estes procariontes, juntamente com eucariontes unicelulares fotossintetizantes, são responsáveis por parte significativa da biomassa encontrada na biosfera (CANNEL, 1993), com algumas estimativas alcançando uma biomassa cianobacteriana total de aproximadamente 10^{15} g (GARCIA-PICHEL et al., 2003). Também estima-se que o gênero *Trichodesmium* (Cyanobacteria, Oscillatoriales) seja responsável por 50% de todo o nitrogênio fixado nos oceanos (BERMAN-FRANK et al., 2003). Além disso, cianobactérias podem ser empregadas na produção de complementos nutricionais, adubos verdes, conversão de energia solar em energia potencial química, e produção de produtos farmacológicos (KREITLOW; MUNDT; LINDEQUIST, 1999; SKULBERG, 1995). Contudo, a capacidade das cianobactérias para a produção de moléculas bioativas também é causa da rica produção de cianotoxinas, as quais são substâncias de uma ampla variedade de bioatividades, e são motivo de preocupação e interesse em todo o mundo (JONES et al., 2009; NUNNERY; MEVERS; GERWICK, 2010).

Desta forma, o metabolismo evidentemente versátil das cianobactérias permite que ocupem nichos ecológicos muito diversos, incluindo ambientes que sofrem perturbações antrópicas. Em alguns casos, tais perturbações podem levar ao aumento demasiado da densidade de células cianobacterianas em um ambiente. Isso ocorre especialmente quando as condições se tornam muito favoráveis para o grupo, ou desfavoráveis para grupos competidores por recursos. Assim, as florações podem ocorrer em corpos d'água doce próximas a ocupações humanas, em áreas industriais, agrárias ou urbanas (DIAS et al., 2004). O contato dos reservatórios e dos corpos d'água com atividades antrópicas resulta, muito frequentemente, na deposição acentuada de substâncias poluentes nitrogenadas e fosforadas, levando a uma situação ótima para a colonização por cianobactérias. Em alguns casos, tais deposições ocorrem sem interferência humana. Nestas situações, a ocorrência de florações cianobacterianas torna-se frequente.

As causas ecológicas das florações não são totalmente entendidas, porém alguns estudos apontam para a possibilidade de serem resultado mais frequente em eventos de perturbações, tanto naturais quanto antrópicas. Em contrapartida, outros estudos apontam para a ocorrência das florações em situações naturais não perturbadas. De qualquer maneira, é

evidente a correlação entre as atividades humanas e o aumento da frequência, da permanência e da intensidade das florações (PAERL; OTTEN, 2013). Estas causam efeitos danosos diversos no ambiente, notoriamente quando frequentes.

Entre os efeitos ecológicos está o aumento da turbidez da água, inibindo o desenvolvimento de outros organismos dependentes da luz solar, em especial os fotossintetizantes. Outro efeito seria a diminuição intensa do conteúdo de oxigênio gasoso dissolvido na água – tornando-o especialmente indisponível aos heterótrofos de tais ecossistemas. Este efeito pode levar à hipoxia e mesmo à anoxia dos ambientes, causando profundos danos ecológicos (NUNNERY; MEVERS; GERWICK, 2010). Um terceiro efeito, de grande preocupação à sociedade, é a liberação de cianotoxinas na água de consumo ou de contato humano ou de animais domésticos. Estas toxinas podem causar uma ampla variedade de efeitos à flora e à fauna, inclusive ao ser humano, causando desde irritações cutâneas até hemorragias internas severas e tumores hepáticos, ou mesmo paralisias musculares concomitantes a asfixiamentos em decorrência de neuropatias agudas (TILLET et al. 2000).

2.2. Metabolismo secundário e vias de síntese de moléculas bioativas em cianobactérias

O metabolismo secundário é todo o metabolismo relacionado a atividades diferentes daquelas consideradas essenciais para uma célula ou organismo, como crescimento, reprodução e manutenção do aparato metabólico em si (NELSON; COX, 2008). Contudo, tal definição carece de conceitos que justifiquem a decisão arbitrária a respeito de qualquer atividade metabólica não totalmente compreendida – seja na perspectiva fisiológica, ecológica ou evolutiva – considerando-a fundamental ou auxiliar, para um organismo, perante as demandas de seu ambiente. Além disso, a definição carece de uma estrutura positiva para incluir quais processos o metabolismo secundário abarca, ao invés de servir somente como um conjunto de elementos arbitrariamente excluídos do metabolismo primário. De qualquer maneira, consideramos o metabolismo primário como aquele conjunto de relações metabólicas bastante conservado entre grupos, conquanto o metabolismo secundário é considerado como variável entre táxon, e em algumas circunstâncias podendo ser variável entre populações ou linhagens de mesma espécie (SASSO et al., 2011).

O filo Cyanobacteria possui uma ampla diversidade na química e na atividade de suas substâncias bioativas, além de estruturas orgânicas incomuns, o que lhes confere um potencial exploratório relevante para obtenção de moléculas de interesse biotecnológico, farmacêutico e medicinal. As possibilidades terapêuticas destas substâncias abrangem

atividades antitumoral, antiviral, antibacteriana, antifúngica, anti-inflamatória, e inibidoras enzimáticas, entre outras (DITTMANN et al., 2015).

São conhecidos atualmente mais de 1100 metabólitos secundários de cianobactérias, dentre as quais 731 foram demonstradas com atividade contra uma ampla variedade de células eucariontes e procariontes, e encontrados em 39 gêneros diferentes. Suas classes estruturais incluem peptídeos, policetídeos, alcaloides, lipídeos, e terpenos, além de híbridos entre essas classes. Mais de dois terços destas moléculas foram encontradas até o momento somente nos gêneros *Microcystis*, *Lyngbya*, *Nostoc* e *Hapalosiphon* (DITTMANN et al., 2015).

Algumas das substâncias bioativas são sintetizadas pela via ribossomal, incluindo algumas classes de peptídeos (cíclicos ou lineares), terpenos, indole-alcaloides, e outros (DITTMANN et al. 2015). Os peptídeos bioativos sintetizados pela via ribossomal têm sua diversidade estrutural resumida aos 20 diferentes aminoácidos proteinogênicos, mas ainda permitindo uma variedade incomum de estruturas orgânicas, como no caso das microviridinas e cianobactinas, devido especialmente a modificações pós-traducionais (LEIKOSKI et al., 2013). As microviridinas e as cianobactinas são substâncias encontradas em aproximadamente 30% dos genomas cianobacterianos sequenciados (DONIA; SCHMIDT, 2011).

Um dos terpenos mais frequentemente encontrados entre aqueles produzidos por bactérias é a geosmina, uma substância volátil de odor característico, e frequentemente associado à produção por actinobactérias, que são os principais produtores desta substância presentes no solo. A geosmina é o produto da atividade enzimática de uma única enzima (geosmina sintase), que realiza a ciclização da molécula farnesil-difosfato (GIGLIO et al., 2008). A produção de geosmina e outros terpenos foi constatada em cianobactérias, em linhagens de espécies do gênero *Nostoc* simbiotes de líquens, além de outras 21 linhagens de vida livre pertencentes a seis gêneros diferentes, que possivelmente representam os principais produtores da molécula em ambientes aquáticos (SUURNÄKKI et al., 2014). A presença de geosmina em corpos d'água potável é responsável por odor e gosto pungentemente desagradáveis, prejudicando tanto o consumo destas águas, quanto o de peixes ali presentes (devido à sua deposição em tecidos lipídicos e musculares de vertebrados), tornando a geosmina uma substância requerente de monitoramento, apesar de não apresentar toxicidade (WATSON et al., 2008).

Outros peptídeos bioativos produzidos por cianobactérias são produzidos pela via não-ribossomal, apresentando uma maior diversidade de monômeros.

A síntese não ribossomal utiliza megassintetases de NRPS (*Nonribosomal Peptide Synthetase*) e/ou PKS (*Polyketide Synthase*), sendo estes complexos enzimáticos codificados

por agrupamentos gênicos, usualmente organizados sob um ou poucos *operons* (WELKER; VON DÖHREN, 2006).

Na síntese utilizando o complexo NRPS, podem ser utilizados mais de 300 tipos de aminoácidos diferentes, inclusive aqueles não proteinogênicos, resultando em diversas possibilidades de estruturas e bioatividades destas moléculas (FINKING; MARAHIEL, 2004; KEHR; PICCHI; DITTMANN, 2011). Os pré-peptídeos do complexo são constituídos por um peptídeo líder e um peptídeo central, associados a enzimas de modificação pós-traducional, as quais catalisam macrociclicações do peptídeo central e outras modificações na cadeia lateral. A finalização do peptídeo é realizada a partir da clivagem catalisada por proteases, e o subsequente transporte através da membrana plasmática.

2.3. Via da Sintetase de Peptídeos Não Ribossomais (NRPS)

As NRPSs possuem organização modular, sendo cada módulo responsável pela incorporação ou modificação do aminoácido na cadeia nascente (MARAHIEL; STACHELHAUS; MOOTZ, 1997; VON DÖHREN et al., 1997). Estes módulos contêm diferentes domínios, classificados segundo sua atividade sobre o substrato.

Entre os módulos encontrados temos o domínio de adenilação (A), responsável pelo reconhecimento e seleção do aminoácido a ser incorporado na cadeia nascente; a proteína carreadora de peptidil (PCP), responsável pelo transporte da cadeia nascente de um centro catalítico para o próximo; domínio de condensação (C), responsável pela catálise da ligação peptídica, considerado então como módulo principal na atividade de alongamento da cadeia polipeptídica; e, por fim, o domínio tioesterase (T), que libera a cadeia do complexo e pode estar envolvida com a ciclização de algumas moléculas.

Ainda, outros domínios auxiliares podem fazer parte da estrutura de uma megassintetase da via NRPS, como domínio metiltransferase (MT), epimerase (E), e outros (WELKER; VON DÖHREN, 2006). Alguns agrupamentos não modulares contendo somente um domínio podem ser encontrados, como no caso descrito por DU & SHEN (1999), onde uma proteína ativa carreadora de peptídeo (BlmI) do agrupamento relacionado à síntese de bleomicina foi encontrada nestas condições.

Fica evidente a necessidade de reconhecimento da ordenação dos genes responsáveis pelo funcionamento apropriado dos módulos da megassintetase. A simples notação da funcionalidade (por homologia de sequência) e de estruturas subgênicas (como sequências de domínios proteicos) é informativa a respeito da presença ou ausência do agrupamento responsável pela síntese de determinada substância pré-definida, além de conhecimento do

potencial funcionamento qualitativo dos peptídeos gerados pelos genes contidos no agrupamento gênico. Contudo, tal notação não é suficiente para indicar se o agrupamento em sua totalidade é funcional, ou ainda se possui algum tipo de região variável em seu produto megassintético. Também, a ordenação dos genes no agrupamento pode indicar processos evolutivos. Desta maneira, modificações na estrutura, assim como na sequência de bases de cada gene do agrupamento, são informações igualmente importantes na análise comparativa.

A síntese não ribossômica realizada por megassintetases de policetídeos é pertencente exclusivamente ao metabolismo secundário (KEHR; PICCHI; DITTMANN, 2011). Assim como as enzimas do complexo NRPS, em algumas instâncias do sistema de biossíntese que utiliza PKS, as enzimas são multifuncionais e modulares, por sua vez utilizando diferentes tipos de ácidos carboxílicos como substrato para a síntese (ROUNGE et al., 2009), em especial a acil-coenzima A e acilmalonil-coenzima A, agindo pela condensação descarboxilativa sequencial dos precursores.

Os módulos (ou subunidades, dependendo da categoria de PKS) são também frequentemente classificados segundo sua atividade sobre o substrato, entre eles o domínio cetoacil-sintase (KS) propriamente dito, que age na etapa de catálise da ligação C-C, domínio aciltransferase (AT), domínio cetorredutase (KR), domínio desidratase (DH), domínio enoil-redutase (ER), e a proteína carreadora de acila (ACP), responsável pela ativação do substrato e pela condução da cadeia nascente policetídica ao longo do aparelho megassintético.

A síntese pela via da PKS pode ser classificada em três categorias (PKS-I, -II e -III) segundo a organização de seus domínios catalíticos. A categoria PKS-I agrupa os aparelhos megaenzimáticos que se valem da estratégia modular similar à organização da NRPS. Desta maneira, cada módulo é constituído por domínios catalíticos responsáveis pelo reconhecimento, ativação e condensação de monômeros, além da eventual modificação da cadeia lateral (FISCHBACH; WALSH, 2006), sendo a síntese de moléculas ocorrendo em um único ciclo não iterativo (SHEN, 2003). Exemplo de uma substância de importância médica biosintetizada por esta via é a eritromicina A. Similar também às NRPSs, as PKS-I possuem características que permitem, dentro de certos limites particulares para cada categoria de metabólito, a predição generalizada da estrutura do produto, desde que sejam observadas não somente as sequências dos genes contidos no agrupamento gênico, mas também seu ordenamento.

Nas categorias PKS-II e PKS-III, os sítios catalíticos são encontrados em peptídeos diferentes. As PKS tipo II são complexos multienzimáticos possuindo um único conjunto de centros catalíticos que agem iterativamente, gerando policetídeos usualmente aromáticos e

policíclicos. Como exemplo de molécula sintetizada por esta categoria, podemos citar a tetracenomicina C. As PKS-III (conhecidas como “similares à calcona-sintase”, calcona sendo uma cetona aromática), são complexos enzimáticos homodiméricos que iterativamente condensam a síntese de substâncias aromáticas, frequentemente monocíclicas ou bicíclicas (SHEN, 2003). Ainda, PKS tipo I e II agem sobre o precursor com a ajuda do domínio de ACP, que conduz a cadeia nascente no aparato enzimático, conquanto as PKSs tipo III agem diretamente sobre o substrato acil-coenzima A, independentemente da ACP.

É evidente, também, que nos casos de PKS tipo II e III, a predizibilidade de produtos metabólicos é mormente prejudicada, pois a organização física da megassintetase não guarda correspondência com sua atividade de maneira tão direta e evidente como ocorre na maioria das NRPS. Ainda, megassintetases PKS guardam similaridade de atividade e sequência com complexos relacionados ao metabolismo de lipídeos, o que torna substancialmente mais complexa a análise de potenciais produtos metabólicos sintetizados por esta via.

2.4. Genômica de cianobactérias

O primeiro sequenciamento genômico procarioto foi finalizado em 1995, de um exemplar da espécie *Haemophilus influenzae* (FLEISCHMANN et al., 1995), enquanto *Synechocystis* sp. PCC 6803 foi a primeira espécie de cianobactéria a ser sequenciada, pouco mais de um ano depois (KANEKO et al., 1996). *Synechocystis* sp. PCC 6803 é uma linhagem de hábito dulcícola, possui até 12 cópias de seu cromossomo (cada cópia contendo 3,57 Mpb) por célula, além de outros 7 plasmídeos. Apesar da relativa precocidade do sequenciamento e anotação dos genomas cianobacterianos em relação às outras bactérias, o acúmulo de sequências de cianobactérias não ocorreu na mesma velocidade do que em outros grupos, como o filo Proteobacteria.

Esta discrepância é evidente quando analisamos o estado atual das informações depositadas no NCBI: quase 80.000 sequências de procariotos, sendo que somente 426 (0,53%) representam genomas de cianobactérias. Contudo, uma parcela relativamente alta de sequências genômicas de cianobactérias apresenta-se na categoria completo (N = 97; 22,77%), apesar das dificuldades inerentes ao seu genoma. Dentre os genomas cianobacterianos desta categoria, o menor sequenciado até o momento é o genoma de *Prochlorococcus* sp. W3, com 0,33 Mpb; conquanto o maior genoma pertence a *Mastigocoleus testarum* BC008, com 15,86 Mpb. Podemos observar, portanto, uma grande variação no tamanho dos genomas cianobacterianos.

Algumas espécies cianobacterianas podem apresentar em seu conteúdo celular mais de uma molécula do mesmo cromossomo, mais de uma molécula do mesmo plasmídeo, ou ainda plasmídeos diferentes, de maneira que a oligoploidia e a poliploidia são fenômenos frequentes no grupo. Em algumas destas espécies poliploides, pode-se encontrar até 450 cópias do mesmo cromossomo em seus acinetos (SUKENIK et al., 2012). Na fase de crescimento exponencial, pode-se encontrar, em alguns casos, mais de 200 cópias do mesmo cromossomo em uma célula não especializada (GRIESE; LANGE; SOPPA, 2011). Estas observações apontam para a necessidade de uma rápida resposta a elicitores ambientais, de forma que, possivelmente, muitos cromossomos são capazes de receber o estímulo e prontamente expressar a resposta. Esse contato entre os cromossomos no interior de uma única célula pode ser, também, causa de mais frequentes recombinações e rearranjos gênicos.

As forças que agem para moldar o cromossomo e resultar em variações no conteúdo genético de procariotos podem ser resumidas, em geral, a eventos de rearranjo e mutações (KRAUSE et al., 2014; BLAJZE et al., 2013). As mutações ocorrem como resultado de erros inerentes ao aparato de replicação da molécula de DNA, assim também como resultado da combinação de influência externa (como agentes mutagênicos), e da sensibilidade do organismo a estas influências. Esta sensibilidade é em decorrência de muitos fatores próprios do organismo, como eficiência dos mecanismos de reparo e produção de substâncias capazes de evitar a incidência ou a atividade de agentes mutagênicos. Algumas mutações são em decorrência de erros no próprio aparato de duplicação gênica, no momento da duplicação da molécula de DNA. Os rearranjos também podem ser resultado de eventos de transferências horizontais, por exemplo através de elementos transponíveis, e recombinações homólogas, sendo produtos de duplicações, inversões, inserções e deleções entre genomas (GRAY, 1999; GRAY, 2000).

O entendimento de rearranjos e de mutações é o cerne dos estudos comparativos sobre divergência entre grupos. Kunin e Ouzonis (2003) apontam que entre a perda e o ganho de função, e a transferência horizontal de genes funcionais em cromossomos procariontes, a perda e o ganho se mostram, em média, em equilíbrio, e são mais significativos para o conteúdo genômico do que a transferência horizontal. Por outro lado, Lin e colegas (2011) mostram que o conteúdo de elementos transponíveis em cianobactérias pode variar entre nenhuma inserção (linhagens *Prochlorococcus* sp. MIT 9211 e MIT 9215), até mais de 500 eventos de inserção em *Microcystis aeruginosa* NIES-843, correspondendo a 10,85% do seu conteúdo genômico total. Esse grande número de inserções também foi observado em outras

linhagens dessa espécie, como *M. aeruginosa* PCC 7806, que apresentou mais de 350 sequências de inserção, representando 8,98% de seu conteúdo genômico.

Estes estudos são representativos sobre a importância entre os fenômenos que definem o conteúdo genômico, e que moldam o cromossomo procarionte, evidenciando a variabilidade destas características entre táxons. As generalizações a respeito são otimistas, no melhor dos casos (HENDRICKSON; LAWRENCE, 2006; OCHMAN; DAVALOS, 2006; KOONIN, 2011). Em especial, podemos observar a frequente incidência destes eventos, e sua influência no conteúdo e na organização genômica da espécie *Microcystis aeruginosa*.

Análises de sequenciamentos genômicos de diversas espécies microbianas, nos últimos anos, têm mostrado que muitos genomas contêm mais genes e agrupamentos gênicos relacionados com a biossíntese de metabólitos secundários do que são conhecidas e isoladas as moléculas bioativas potencialmente produzidas por estas espécies ou linhagens (WINTER et al., 2011).

Isto também evidencia a necessidade de aplicação de recursos no estudo de sequenciamentos já disponíveis, pois sequenciamentos de linhagens ainda não realizados são frequentemente favorecidos em detrimento de aprofundamentos a respeito de informações valiosas ainda por serem explorados. A exemplo, genomas de táxons intensamente estudados, como *Streptomyces* e *Aspergillus*, conhecidas pela vasta diversidade de metabólitos, têm fornecido informações abundantes sobre moléculas ainda não descobertas (SANCHEZ et al., 2008).

Possivelmente, o fator principal para esta discrepância seja o fato de muitos desses metabólitos secundários serem produzidos sob circunstâncias específicas no meio natural, assim os genes de muitos táxons menos estudados permanecem silenciosos nas condições de laboratório, onde elicitores específicos, e desconhecidos, não ocorrem. É possível que uma das maiores barreiras para descoberta de novas substâncias bioativas seja o desconhecimento das circunstâncias necessárias para a atividade de tais agrupamentos (SASSO et al., 2012).

3. HIPÓTESES#

H₀: A hipótese nula deste trabalho propõe que as características genômicas relacionadas ao metabolismo secundário são iguais entre todas as linhagens de *Microcystis aeruginosa*. Portanto, espera-se que a quantidade e organização de agrupamentos, assim como a quantidade de categorias estruturais diferentes de metabólitos potencialmente expressos, sejam iguais entre todas as linhagens analisadas;

H_A: A hipótese alternativa propõe que as características genômicas relacionadas ao metabolismo secundário são diferentes entre as linhagens de *M. aeruginosa*. Desta maneira, espera-se que a quantidade e organização de agrupamentos, assim como a quantidade de categorias estruturais de metabólitos, sejam diferentes entre as linhagens analisadas, mostrando variabilidade nas características genômicas relacionadas ao metabolismo secundário.

#

4. OBJETIVOS#

O objetivo geral deste estudo foi comparar genomicamente 18 linhagens da espécie *Microcystis aeruginosa*, com ênfase no seu metabolismo secundário.

Os objetivos específicos foram:

- I. Encontrar padrões genômicos relevantes à compreensão da espécie;
- II. Analisar as possíveis variações nas quantidades, categorias e estruturas de agrupamentos gênicos relacionados à produção de moléculas bioativas na espécie, em especial aquelas originadas da síntese não ribossomal;
- III. Inferir relações estatísticas entre as características genômicas de interesse, em especial as que se referem à síntese de metabólitos secundários, permitindo sugerir processos que levam aos padrões genômicos observados.

5. MATERIAIS E MÉTODOS#

5.1. Origem das linhagens

Para desenvolvimento deste trabalho foram utilizadas 18 sequências do genômicas da espécie *Microcystis aeruginosa*, disponíveis no banco de dados GenBank, do National Center for Biotechnology Information (NCBI), de diversos continentes (Tabela 1).

Tabela 1 – Linhagens utilizadas para a comparação genômica, e origem geográfica

Número de acesso	Quantidade Seq. Contíguas	Linhagem	Origem
NZ_ASZQ000000000.1	278	SPC777	Represa Billings, São Paulo - Brasil
NZ_AQPY000000000.1	1.363	PCC 7005	Lago Mendota – EUA
NZ_CAIQ000000000.1	303	PCC 9809	Lago Mishigan – EUA
NZ_CAIL000000000.1	93	PCC 9806	Oshkosh – EUA
NZ_CAIH000000000.1	132	PCC 9432	Lago Little Rideau - Canadá
NZ_CAIK000000000.1	77	PCC 7941	Lago Little Rideau - Canadá
NZ_CAI000000000.1	264	PCC 9717	Represa Rochereau - França
NZ_CAIQ000000000.1	323	PCC 9701	Represa Guerlesquin - França
NZ_CAIJ000000000.1	221	PCC 9443	Lagoa de pescaria – República Central Africana
NZ_CAIM000000000.1	267	PCC 9807	Represa Hartbeespoort – África do Sul
NZ_CAIN000000000.1	141	PCC 9808	Represa Malpas - Austrália
NZ_AOCI000000000.1	355	DIANCHI905	Lago Dianchi – China
NZ_ANQK000000000.1	4	TAIHU98	Lago Taihu – China
NZ_JXYX000000000.1	262	NIES-88	Lago Kasumigaura Ibaraki - Japão
NZ_BBPA000000000.1	79	NIES-44	Lago Kasumigaura Ibaraki – Japão
NZ_CP012375.1	1	NIES-2481	Lago Kasumigaura Ibaraki – Japão
NZ_CP011304.1	1	NIES-2549	Lago Kasumigaura Ibaraki – Japão
NC_010296.1	1	NIES-843	Lago Kasumigaura Ibaraki – Japão

5.2. Localização e anotação dos agrupamentos relacionados com o metabolismo secundário, análises genômicas e estatísticas

A localização dos agrupamentos gênicos relacionados ao metabolismo secundário foi realizada utilizando a plataforma antiSMASH 3.0 (MEDEMA et al., 2011). Esta plataforma realiza a procura com sequências-assinatura referenciais em genomas. Seu banco de dados é

atualizado com frequência, permitindo que a plataforma localize agrupamentos relacionados a metabólitos secundários mais recentemente descobertos, aumentando a confiabilidade das análises. Quando necessário, os programas BLAST, Pfam e Gview foram utilizados para a elucidação de genes de maneira mais pormenorizada, ou para o entendimento da organização dos agrupamentos.

O programa Artemis foi utilizado para a anotação manual de agrupamentos de maior interesse, utilizando o banco de dados Minimum Information about a Biosynthetic Gene Cluster (MIBiG) para a obtenção de padrões para classificação e alinhamento de agrupamentos gênicos envolvidos com a biossíntese de moléculas bioativas.

A fim de verificar a organização relativa do genoma entre os 18 utilizados neste trabalho, o programa Mauve (DARLING et al., 2004) foi utilizado. O algoritmo progressiveMauve (DARLING; MAU; PERNA, 2010) é capaz de alinhar regiões de sequência sintênicas entre dois ou mais genomas, na tentativa de encontrar blocos sintênicos colineares, ou a ausência destes.

Considerando a complexidade de organização genômica observada entre os genomas da espécie, estudos sobre a homologia da sequência completa do genoma se mostram interessantes, de forma que permitem compreender como a espécie se estrutura genomicamente, o que é conservado entre todos os genomas utilizados neste estudo, além do que pode ser peculiarmente variável. Para a visualização da similaridade de todos os genomas das linhagens analisadas, foi utilizado o software BRIG-0.95 (ALIKHAN et al., 2011), capaz de gerar, dentre uma ampla gama de opções, atlas circulares que abrangem um ou mais genomas, alinhando localmente sua(s) sequência(s) com relação a um genoma de referência.

Esta ferramenta é eficaz para a detecção visual de regiões incomuns que podem sugerir algum processo de significado biológico – como, por exemplo, regiões de baixa similaridade, as quais podem indicar regiões potenciais de inserção, transferência horizontal, ou variabilidade. Em contrapartida, esta ferramenta também é capaz de mostrar regiões altamente conservadas.

Com o objetivo de definir a totalidade de genes das linhagens deste estudo, um pangenoma foi quantitativamente estimado a partir das sequências de genes, de forma que a comunalidade entre as linhagens que define a espécie, assim como o grau de conhecimento a respeito das diferenças entre as linhagens, pudesse ser melhor averiguada. Para tanto, o software GET_HOMOLOGUES (CONTRERAS-MOREIRA. VINUESA, 2013) foi utilizado. Este programa produz uma matriz de similaridade entre genes de sequências fornecidas, desta

forma permitindo algumas estimativas, entre elas a totalidade de homólogos entre pelo menos dois genomas, homólogos entre todos os genomas, e genes presentes em somente um genoma.

O software GET_HOMOLOGUES também foi utilizado para a inferência do genoma cerne (coregenome). A análise procede de maneira similar à inferência do pangenoma, contudo o que se avalia são os genes homólogos (>85% de identidade, > 75% de cobertura) comuns a todas as linhagens ou genomas em análise.

Utilizando-se métodos estatísticos descritivos e inferenciais, foram realizadas análises a partir dos 18 genomas selecionados para este estudo, objetivando-se um panorama quantitativo da espécie. Para tais análises, foram utilizados seis diferentes variáveis, sendo elas: i) quantidade de transposases preditas em cada genoma; ii) tamanho, em megapares de bases, de cada genoma; iii) quantidade de genes preditos para cada genoma; iv) quantidade de agrupamentos gênicos relacionados a metabólitos secundários preditos para cada genoma; v) quantidade de sequências contíguas para cada genoma; e vi) quantidade de categorias de metabólitos preditos genomicamente.

A variável quantidade de transposases preditas se mostra interessante devido ao seu potencial como indicador de padrões e intensidades da dinâmica genômica das linhagens e da espécie. Considerando que a presença destas sequências indica uma possível ocorrência de inserções ou de transferência horizontal, eventos os quais podem também estar relacionados a inversões e deleções, além da própria mudança na sequência genômica em blocos ou regiões.

A variável tamanho de cada genoma foi escolhida devido a sua potencial relação com outras variáveis utilizadas para caracterizar a amostra, como quantidade de transposases preditas, genes preditos e quantidade de agrupamentos preditos. A motivação desta escolha se baseia na correspondência aproximada da relação entre informação e pares de base, que nos leva a supor que o tamanho do genoma procariótico é aproximadamente proporcional à riqueza de informações ali contidas. Além do mais, um dos objetivos principais deste trabalho é a avaliação e comparação da variedade e da frequência dos agrupamentos gênicos relacionados a metabólitos secundários, fatores que poderiam ser correlacionados ao tamanho dos genomas.

A variável quantidade de genes preditos para cada genoma foi utilizada com o objetivo de se levantar padrões mais diretamente associados a características genômicas funcionais. Ainda na mesma direção do raciocínio adotado para a variável tamanho de cada genoma, supomos uma relação entre a quantidade de genes preditos com o potencial para se encontrar mais agrupamentos relacionados ao metabolismo secundário.

A variável quantidade de agrupamentos gênicos relacionados a metabólitos secundários preditos para cada genoma é a variável com proximidade mais direta em relação aos objetivos principais do trabalho, no qual buscou-se comparar a riqueza e estrutura dos agrupamentos relacionados à síntese de moléculas bioativas. Desta maneira, esta variável é central para o trabalho, pois caracteriza direta (através do levantamento e análise da variável) e indiretamente (através das correlações com as outras cinco variáveis) o sistema gênico na amostra, e tenta extrapolar tais informações para a espécie.

A variável quantidade de sequências contíguas definidas no sequenciamento disponível para cada genoma representa a interferência dos métodos genômicos nos estudos comparativos. Considerando que uma das maiores dificuldade apresentadas nos procedimentos genômicos de genomas cianobacterianos é a etapa de montagem, é interessante compreender quantitativamente como os processos de sequenciamento e montagem podem oferecer vieses ou artefatos no momento de análises comparativas. Portanto, a análise desta variável proporciona (notoriamente em suas correlações com as outras variáveis) a perspectiva biológica do genoma e seu conteúdo sob a influência destes interferentes dos métodos genômicos.

A última variável considerada para a caracterização genômica da espécie *Microcystis aeruginosa* foi a quantidade de categorias de metabólitos preditos genomicamente. Servindo como um complemento à variável quantidade de agrupamentos gênicos relacionados a metabólitos secundário preditos, pois esta última abrange quantitativamente, porém carece de expressão a respeito da diversidade de metabólitos secundários potencialmente sintetizados pela espécie, e sua variação entre as linhagens. A variável quantidade de categorias supre esta carência, e torna as análises principais deste trabalho melhor caracterizadas.

Estas variáveis foram submetidas ao teste Shapiro-Wilk de normalidade, de maneira a ser averiguada sua aderência à distribuição normal de dados, permitindo que a estatística inferencial possa corroborar de maneira realista as extrapolações da amostragem para a espécie. O p-valor considerado para esta análise foi $p < 0,01$, pois a amostragem é claramente pequena, demandando um relaxamento da limite de aceitação. Desta maneira, valores maiores do que 0,01 foram considerados de frequência normalmente distribuída, e aptos para as análises de correlação linear, bem como para as inferências populacionais.

A partir dessas variáveis, foram realizadas 15 análises de coeficiente de correlação linear de Pearson, com o objetivo de procurar por variáveis que se associem causativamente, na tentativa de explicar alguns padrões encontrados nas estatísticas descritivas e estimativas da população a partir da amostra (Tabela 2).

Tabela 2 – Indicação dos pares de variáveis avaliadas em cada correlação

Primeira Variável	Segunda Variável
Quantidade de transposases preditas em cada genoma	Tamanho de cada genoma em megapares de bases
Quantidade de transposases preditas em cada genoma	Quantidade de genes preditos para cada genoma
Quantidade de transposases preditas em cada genoma	Quantidade de agrupamentos gênicos relacionados a metabólitos secundários preditos
Quantidade de transposases preditas em cada genoma	Quantidade de sequências contíguas para cada genoma
Quantidade de transposases preditas em cada genoma	Quantidade de categorias de metabólitos preditos genomicamente
Tamanho de cada genoma em megapares de bases	Quantidade de genes preditos para cada genoma
Tamanho de cada genoma em megapares de bases	Quantidade de agrupamentos gênicos relacionados a metabólitos secundários preditos
Tamanho de cada genoma em megapares de bases	Quantidade de sequências contíguas para cada genoma
Tamanho de cada genoma em megapares de bases	Quantidade de categorias de metabólitos preditos genomicamente
Quantidade de genes preditos para cada genoma	Quantidade de agrupamentos gênicos relacionados a metabólitos secundários preditos
Quantidade de genes preditos para cada genoma	Quantidade de sequências contíguas para cada genoma
Quantidade de genes preditos para cada genoma	Quantidade de categorias de metabólitos preditos genomicamente
Quantidade de agrupamentos gênicos relacionados a metabólitos secundários preditos	Quantidade de sequências contíguas para cada genoma
Quantidade de agrupamentos gênicos relacionados a metabólitos secundários preditos	Quantidade de categorias de metabólitos preditos genomicamente
Quantidade de sequências contíguas para cada genoma	Quantidade de categorias de metabólitos preditos genomicamente

Com o objetivo de verificar o comportamento da quantidade de sequências contíguas na contagem de agrupamentos gênicos em genomas não completos, uma simulação foi realizada. Nesta simulação, o genoma das três linhagens de genoma completo (NIES-843, NIES-2481 e NIES-2549) foram fragmentados computacionalmente. Foram gerados entre 10 e 2.000 fragmentos para cada genoma, a intervalos de 10 fragmentos. Estes fragmentos de tamanho regular foram analisados com o uso da plataforma antiSMASH 3.0 para a localização dos agrupamentos, e então a contagem foi plotada e comparada entre as três linhagens.

6. RESULTADOS E DISCUSSÃO#

6.1. Agrupamentos gênicos relacionados a produtos naturais encontrados na espécie *Microcystis aeruginosa*#

As 18 linhagens de *Microcystis aeruginosa* foram submetidas à análise na plataforma antiSMASH, obtendo-se um panorama da diversidade de tipos de agrupamentos gênicos relacionados ao metabolismo secundário da espécie, além de indicativos gerais sobre organização e variações nos agrupamentos.

Esta análise permitiu a localização de 235 agrupamentos gênicos relacionados à biossíntese de moléculas bioativas nas 18 linhagens, categorizadas em 12 classes segundo as estruturas de seus produtos. Essa grande variedade de agrupamentos e de classes estruturais é expressão da riqueza de moléculas bioativas produzidas por esta espécie cianobacteriana. GRAM (2014) relata a descoberta de 15 agrupamentos em *Pseudoalteromonas luteoviolaceae*, espécie pertencente a um gênero reconhecido pela diversidade incomum de moléculas bioativas. WILSON et al. (2014), valendo-se de técnicas de metagenômica, mostraram a presença de pouco mais de 40 agrupamentos diferentes, divididos em 7 classes estruturais, em um único filo (filo candidato Tectomicrobia). O filo Cyanobacteria é conhecido por sua variedade de produtos naturais estruturalmente incomuns, sendo o gênero *Microcystis* abundante na síntese de moléculas bioativas: mais de 250 tipos diferentes de metabólitos secundários já foram descritos somente para este gênero (DITTMANN et al., 2015).

Das 12 classes estruturais encontradas na espécie *M. aeruginosa*, as classes com maior número de agrupamentos foram os terpenos (53; 22,55%), os híbridos (50; 21,27%), as bacteriocinas (32; 13,61%), e as NRPSs (29; 12,34%), perfazendo um total de 164 agrupamentos gênicos (69,78%). As outras classes, com menor número de agrupamentos, foram as PKSs tipo I (24; 10,21%), cianobactinas (16; 6,8%), microviridinas (11; 4,68%), PKSs tipo III (5; 2,12%), resorcinol (2; 0,85%), lassopeptídeos (1; 0,42%). Também foram encontrados agrupamentos contendo uma sintase de policetídeo potencialmente associado ao metabolismo lipídico, porém que não pôde ser classificada e identificada mais especificamente (6; 2,55%), e uma categoria abrangendo tipos diferentes de metabólitos que não puderam ser classificados (6; 2,55%) (Figura 1).

Figura 1: Abundância total e relativa de agrupamentos gênicos contidos em genomas de *M. aeruginosa*, de cada classe estrutural de metabólitos secundários, encontrados em todas as linhagens. Nota-se que as quatro classes mais abundantes perfazem aproximadamente dois terços do total de agrupamentos gênicos.

Fonte: Weiss (2017)

6.2. Comparações do metabolismo secundário entre as linhagens

6.2.1. Variação dos agrupamentos relacionados a metabólitos secundários mais frequentemente encontrados nos genomas#

A distribuição dos agrupamentos de metabólitos secundários foi substancialmente diferente entre as linhagens. Em uma das linhagens foram encontrado somente 6 agrupamentos de 4 categorias de metabólitos diferentes (PCC 9806), enquanto em outra observou-se 21 diferentes agrupamentos de 10 categorias de metabólitos (PCC 7005), ainda que em média tenham sido encontrados 13,05 ($\pm 3,50$; limite de confiança de 95%) agrupamentos, e 6,55 ($\pm 1,65$; limite de confiança de 95%) categorias de agrupamentos de moléculas bioativas por linhagem. Estas observações podem ser reflexo da dinâmica incomum do genoma da espécie *Microcystis aeruginosa*, o qual sofre inserções e rearranjos com frequência (LIN et al. 2011). Sobre tudo, é evidente que a espécie apresenta uma alta riqueza de agrupamentos relacionados ao metabolismo secundário.

Somente agrupamentos relacionados a bacteriocinas, terpenos e agrupamentos híbridos foram encontrados em todas as linhagens. Contudo, a classe de híbridos agrupa uma ampla variedade de classes estruturais, devido a sua natureza combinatória. Além disso, o encontro frequente desta classe nos genomas pode ser, em parte, consequência de um viés do método utilizado pela plataforma antiSMASH. Este método consiste em contrastar sequências-assinatura de seu banco de dados contra a sequência-alvo ou genoma, e classificar

os agrupamentos encontrados. Para cada classe estrutural encontrada, está definindo um limite de distância (determinado experimentalmente) a partir das extremidades da sequência-assinatura em que são incluídos os genes no agrupamento para a extrapolação da categoria de metabólito a que o agrupamento refere-se, os quais apresentam um limite mínimo de distância para a identificação (WEBER et al., 2015). Desta forma, existe a possibilidade de serem encontrados genes pertencentes a outro agrupamento dentro da distância pré-definida, levando a plataforma a classificar estes dois agrupamentos separados como um único híbrido. Em alguns casos, isso pode gerar falsos positivos de agrupamentos híbridos e subestimar o número de agrupamentos presentes em um genoma, além de distorcer as categorias de metabólitos secundários que são produto destes agrupamentos. Desta forma, análises detalhadas através do uso de softwares como Artemis, apesar de trabalhosas e demandar tempo para seu aprimoramento, podem colaborar nos estudos de agrupamentos.

Alguns agrupamentos gênicos relacionados a metabólitos secundários foram encontrados em comum a algumas linhagens. Com relação à linhagem brasileira SPC777, uma seleção de agrupamentos foi comparada com aqueles encontrados em outras linhagens, de forma a definir um perfil genômico pormenorizado do sistema de produção destas moléculas bioativas em comum. Três destes agrupamentos gênicos, da categoria NRPS ou híbridos incluído NRPS, foram analisados e comparados, sendo estes: aeruginosina, microcistina e micropeptina. Estes agrupamentos representam os mais frequentemente encontrados destas categorias em todas as linhagens.

6.2.2. Aeruginosina

A comparação da estrutura do agrupamento da aeruginosina foi realizada utilizando o padrão do agrupamento da linhagem *Microcystis aeruginosa* NIES-98, obtido do banco de dados de metabólitos secundários MIBiG. Das 18 linhagens em estudo, 15 apresentaram agrupamento gênico da biossíntese da megassintetase responsável pela síntese deste peptídeo bioativo. Estas 15 linhagens mostram regiões conservadas dentro do agrupamento, assim como regiões de menor conservação. Contudo, somente a linhagem TAIHU98 apresenta todos os genes encontrados no agrupamento padrão, encontrado na linhagem NIES-98 (Figura 7).

O gene *aerJ*, responsável pela função de halogenação do peptídeo, está presente somente em quatro linhagens (NIES-98, NIES-2549, TAIHU98 e DIANCHI905), indicando que nas outras linhagens a halogenação não ocorra, ou ocorra por um processo externo ao agrupamento. O gene *aerG2*, de função desconhecida, porém contendo domínios C e PCP, somente foi encontrado nas linhagens NIES-98, TAIHU98 e PCC 7005, enquanto parte do

gene *aerB*, responsável pela condensação e epimerização da isoleucina no peptídeo nascente, está ausente na maioria das linhagens, com exceção das NIES-98, PCC 9807 e TAIHU98. A linhagem PCC 9807 não apresentou o gene *aerN*, responsável pelo transporte transmembrana do peptídeo. A região central do agrupamento (genes *aerK*, *aerD*, *aerE*, *aerF*, e a região terminal do gene *aerB*) se mostra conservada em todas as linhagens. A linhagem SPC777 mostrou a ausência do gene *aerJ* e *aerG2*, além de baixa identidade dos genes *aerA*, *aerB*, e *aerG1* (Figura 2).

Figura 2 – Agrupamentos gênicos preditos como envolvidos na biossíntese de aeruginosina de cada linhagem. A referência (NIES-98) está em cor cinza, sendo a primeira, de baixo para cima. A sequência das linhagens é, de baixo para cima: NIES-843, NIES-2549, NIES-2481, NIES-88, TAIHU98, DIANCHI905, PCC 9808, PCC 9807, PCC 9443, PCC 9717, PCC 7941, PCC 9432, PCC 9809, PCC 7005 e SPC777. Foi utilizado um padrão de cores específico para a designação do continente de origem de cada linhagem, sendo ele: roxo = Ásia; rosa = Oceania; vermelho = África; amarelo = Europa; verde = América do Norte; e azul = América do Sul.
Fonte: Weiss (2017)

A aeruginosina é um peptídeo não-ribossomal NRPS, inibidor de serina-proteases, especialmente trombina. Sua estrutura pode variar substancialmente em sua região N-terminal, onde algumas formas apresentam ácidos graxos, L-tirosina, L-arginina, ou açúcares pentoses, alterando sua toxicidade e bioatividade (ISHIDA et al., 2009). Ainda, aeruginosinas congêneres podem apresentar diversidade estrutural pela adição de heteroátomos, resultado de reações de halogenação (CADEL-SIX et al., 2008). Desta forma, a classe das aeruginosinas reúne ampla variedade de peptídeos bioativos, sendo as estruturas Hpla (4-hidroxifenil ácido láctico), Choi (2-carboxi-6-hidroxi-octahidroindole), e um derivativo de arginina ou isoleucina, as regiões comuns frequentemente encontradas entre as aeruginosinas – desta maneira os genes associados com a biossíntese destas regiões caracterizam o agrupamento gênico desta molécula bioativa.

Estas análises evidenciam a alta conservação de algumas regiões para a síntese apropriada da molécula bioativa, ao passo que algumas regiões consideradas importantes mostram muito menor conservação de sequência entre as linhagens, evidenciado pela baixa identidade com relação à referência. Isso sugere uma menor relevância destas regiões para a biossíntese de aeruginosina, ou, ainda, um rico potencial da espécie *M. aeruginosa* em produzir variantes desta molécula.

6.2.3. Microcistina

Na comparação da estrutura do agrupamento da microcistina, foi utilizado como padrão o agrupamento da linhagem *Microcystis aeruginosa* PCC 7806, obtido do banco de dados de metabólitos secundários MIBiG. Das 18 linhagens em estudo, apenas 11 apresentaram o agrupamento gênico para biossíntese de microcistina. Nestas linhagens, é possível observar alta conservação da organização e da sequência no agrupamento deste peptídeo bioativo (Figura 8), o que sugere que, naquelas linhagens em que o agrupamento está presente, esta molécula bioativa é potencialmente importante para funções ecológicas, e, portanto, sua sequência é conservada evolutivamente.

Na linhagem NIES-2549, um agrupamento gênico de microcistina foi acusado pela plataforma antiSMASH; contudo, análises mais aprofundadas mostram a presença de somente 3 genes flanqueadores do agrupamento - *dnaN* (beta subunidade DNA polimerase III, a qual é algumas vezes utilizada como um dos marcadores da presença deste agrupamento gênico em *Microcystis aeruginosa*), *uma1* (de função desconhecida, menos utilizado como um marcador da presença do agrupamento) e *uma2* (função desconhecida). Desta forma, a linhagem NIES-2549 apresenta poucos genes do agrupamento, insuficientes para a síntese do metabólito, apesar de manter genes flanqueadores característicos deste agrupamento em *Microcystis*.

A linhagem PCC 9717 apresentou apenas regiões parciais com identidade suficiente nos genes *mcyA*, *mcyB* e *mcyC*, também mostrando uma conformação possivelmente insuficiente para a biossíntese deste peptídeo. As linhagens PCC 9443, PCC 9717, NIES-843 e SPC777 mostram uma região dissimilar no gene *mcyB*, responsável pela condensação de leucina e de ^DN-2-metilasparagina no peptídeo nascente, indicando um potencial para síntese de outras variantes. Ainda, as linhagens PCC 9808 e PCC 9809 apresentam uma extensão desta região dissimilar, além de outra região deste tipo no gene *mcyC* (Figura 3). Estas regiões maiores entre e dentro das regiões gênicas podem resultar em biossíntese de peptídeos diferentes, ou mesmo inabilitar o processo de síntese.

Figura 3 – Agrupamentos potencialmente envolvidos com a biossíntese de microcistina de cada linhagem. A referência (PCC 7806) está em cor cinza, sendo a primeira, de baixo para cima. A sequência das linhagens é, de baixo para cima: NIES-843, NIES-2549, NIES-88, DIANCHI905, PCC 9808, PCC 9807, PCC 9443, PCC 9717, PCC 7941, PCC 9809, e SPC777. Foi utilizado um padrão de cores específico para a designação do continente de origem de cada linhagem, sendo ele: roxo = Ásia; rosa = Oceania; vermelho = África; amarelo = Europa; verde = América do Norte; e azul = América do Sul.

Fonte: Weiss (2017)

A microcistina é um peptídeo bioativo sintetizado por uma via híbrida NRPS/PKS, inibidor de proteínas serina/treonina fosfatase 1 e 2A, sendo responsável por casos conhecidos de toxicidade humana a partir de florações cianobacterianas. No ser humano, a microcistina é absorvida pelo sistema de transporte de ânions, alcançando o fígado. Ali, o metabólito leva as células hepáticas à necrose, causando hemorragias internas severas. A dosagem subletal do peptídeo tem sido apontada como fator causador de desenvolvimento de carcinomas hepáticos primários (TILLET et al. 2000).

As microcistinas são um grupo bastante diverso de heptapeptídeos cíclicos, e compartilham a estrutura $\text{Adda-}^{\text{D}}\text{Glu-Mdha-}^{\text{D}}\text{Ala-}^{\text{L}}\text{X-}^{\text{D}}\text{MeAsp-}^{\text{L}}\text{Z}$, onde X e Z são aminoácidos variáveis, Adda é ácido 3-amino-9-metoxi-2,6,8-trimetil-10-fenil-4,6-decadienoico, $^{\text{D}}\text{MeAsp}$ é ácido 3-metilaspártico, e Mdha é N-metildehidroalanina (TILLET et al., 2000). Esta estrutura é utilizada para a localização e caracterização dos agrupamentos gênicos de biossíntese deste metabólito.

A atividade danosa desta molécula, e sua complexidade estrutural, requerem análises detalhadas para sua compreensão. A atividade de algumas NRPS/PKS, como a microcistina, permite que algumas regiões da estrutura do produto sintetizado sejam variáveis, e dependam tanto da abundância dos aminoácidos disponíveis para a síntese, quanto da própria modificação das sequências codificadoras da megassintetase. Isso permite que uma ampla quantidade de variantes seja possível de biossíntese a partir da estrutura básica do agrupamento da microcistina.

6.2.4. Micropeptina

Para a comparação da estrutura e sequência do agrupamento da micropeptina, foi utilizado o padrão do agrupamento da linhagem *Microcystis aeruginosa* K-139, obtido do banco de dados de metabólitos secundários MIBiG. Das 18 linhagens em estudo, 15 apresentaram o agrupamento gênico para biossíntese de micropeptina. Em algumas linhagens, o agrupamento deste metabólito mostrou-se sintênico. Em contrapartida, outras linhagens mostraram o agrupamento pouco similar ao agrupamento de referência (Tabela 3). Nas linhagens SPC777, TAIHU98 e PCC 9806, o agrupamento (deste ponto em diante chamado de “potencial micropeptina”) mostrou-se estruturalmente mais similar entre si, do que com relação à referência e às outras linhagens. Isso levou a supor a potencialidade de síntese de variantes de micropeptina semelhantes entre essas linhagens e distintas das produzidas pelas restantes, a partir da ação das megassintetases codificadas por este agrupamento.

Todavia, devido à baixa similaridade com a referência, é razoável supor que esse agrupamento codifique não apenas variantes distintas, mas um metabólito substancialmente diferente (Figura 4, Tabela 4), ou mesmo não possua capacidade de síntese.

Tabela 3 – Similaridade informada pela plataforma antiSMASH de cada agrupamento identificado como codificador de enzimas da via biossintética da micropeptina, nas 15 linhagens que apresentaram este agrupamento

Linhagem	Similaridade ao agrupamento de micropeptina
SPC777	37%
PCC 9809	37%
PCC 9806	22%
PCC 9432	87%
PCC 9717	37%
PCC 9701	62%
PCC 9443	50%
PCC 9807	50%
PCC 9808	100%
DIANCHI905	85%
TAIHU98	62%
NIES-88	75%
NIES-2481	87%
NIES-2549	87%
NIES-843	100%

Figura 4 – Organização do agrupamento potencial micropeptina encontrado nas linhagens SPC777, TAIHU98 e PCC9806. Os genes estão listados na Tabela 4, da esquerda para a direita. Em vermelho, genes potencialmente envolvidos diretamente na biossintese do metabólito, possivelmente codificando NRPS ou PKS. Em azul, um gene cujo produto é um transportador ABC (similar a *mcnF*)

Fonte: Weiss (2017)

Tabela 4 – Identidade dos genes encontrados no agrupamento potencial micropeptina encontrado na linhagem SPC777. O padrão é similar ao agrupamento encontrado nas linhagens TAIHU98 e PCC 9806. A ordem se refere à Figura 4, da esquerda para a direita.

Gene	Posição	Senso	Identificação
MAESPC01395	1	+	D-inositol-3-fosfato glicosiltransferase
MAESPC01396	2	-	DNA Metiltransferase
MAESPC01397	3	-	-
MAESPC01398	4	+	6-fosfogluconate desidrogenase
MAESPC01399	5	-	-
MAESPC01400	6	-	-
MAESPC01401	7	-	-
MAESPC01402	8	-	Transportador ABC transmembrana : similar a <i>McyF</i>
MAESPC01403	9	-	Domínio PKS (tioester redutase)
MAESPC01404	10	-	O-Metiltransferase
MAESPC01405	11	-	-
MAESPC01406	12	-	piruvate aminotransferase : similar a <i>McnB</i>
MAESPC01407	13	-	Domínios C – PKSfosfopanteteína – nMT – A : similar a <i>Puwa</i>
MAESPC01408	14	-	Undecaprenil pirofosfate sintase
MAESPC01409	15	-	O-metiltransferase
MAESPC01410	16	-	Corismato-piruvato liase
MAESPC01411	17	-	Colina-sulfatase
MAESPC01412	18	+	-
MAESPC01413	19	+	Sintase dependente de AMP
MAESPC01414	20	+	Proteína de ligação de fosfopanteteína
MAESPC01415	21	+	Oxidase coproporfirinogena-III
MAESPC01416	22	+	AMP sintase-ligase
MAESPC01417	23	+	Gramicidin linear D : Domínio ACP
MAESPC01418	24	-	Gramicidin linear D sintase
MAESPC01419	25	-	-
MAESPC01420	26	-	- : similar a <i>McnC</i>
MAESPC01421	27	-	<i>hetM</i>
MAESPC01422	28	-	-
MAESPC01423	29	-	Domínio sulfotransferase
MAESPC01424	30	-	Acetolactato sintase
MAESPC01425	31	-	Sulfotransferase
MAESPC01426	32	+	- DDE superfamily endonuclease
MAESPC01427	33	+	- DDE superfamily endonuclease

O agrupamento potencial micropeptina encontrado nas linhagens SPC777, TAIHU98 e PCC9806, é constituído por 33 genes preditos (Tabela 4), no tamanho total de aproximadamente 45 kpb, apresentando similaridade em sua organização com um agrupamento encontrado na linhagem PCC 7806 (não incluída nas análises deste estudo), os quais são identificados como micropeptina pela plataforma antiSMASH (Tabela 3).

No agrupamento potencial micropeptina é possível encontrar três módulos principais de adenilação, dos quais um é possivelmente responsável pela incorporação de arginina na cadeia nascente, outro possivelmente responsável pela incorporação de glutamina, cisteína ou isovalina, e o último módulo possivelmente responsável pela incorporação de arginina ou fenilalanina.

Ainda, destes três módulos, somente o primeiro e o último (que apresentaram predições de afinidade a arginina e arginina/fenilalanina) possuem domínios de condensação, regiões necessárias para a devida incorporação de aminoácidos na cadeia em formação. Encontramos também genes similares a hidrolases dependentes de zinco, transposases, lactamases, transportadores ABC, e outros genes ou regiões, inclusive sequências de NRPS e PKS que nos permitem prever parcialmente estruturas para este potencial metabólito (Figura 5).

Figura 5 – Predição da estrutura do metabólito potencialmente sintetizado a partir do agrupamento potencial micropeptina encontrado nas linhagens SPC777, TAIHU98, PCC 9806 e PCC 7806. Predição gerada pela plataforma antiSMASH.

Fonte: Weiss (2017)

Maiores esforços são necessários para elucidar apropriadamente se este agrupamento é um novo agrupamento similar à micropeptina, ou se consiste em uma região sem capacidade codificadora de megassintetases. Outros métodos devem ser adotados, baseados em procedimentos microbiológicos, bioquímicos e de biologia molecular, para a melhor definição destas hipóteses acerca desta região. Contudo, as informações obtidas com as ferramentas

computacionais indicam fortemente que esta região não se trata de um agrupamento funcional.

As linhagens contendo agrupamentos gênicos mais se assemelham à referência são NIES-843, NIES-2549, NIES-2481, DIANCHI98, PCC 9808, e PCC 9432 (Tabela 3). Ainda que mostrem maior similaridade informada pela plataforma antiSMASH, muitas regiões dissimilares são encontradas dentro do agrupamento, e muitos dos genes não possuem identidade suficiente em relação à referência, o que indica fortemente a impossibilidade de síntese desse metabólito (Figura 6).

Figura 6 – Agrupamentos relacionados à micropeptina de cada linhagem. A referência (K-139) está em cor cinza, sendo a primeira, de baixo para cima. A sequência das linhagens é, de baixo para cima: NIES-843, NIES-2549, NIES-2481, NIES-88, TAIHU98, DIANCHI905, PCC 9808, PCC 9807, PCC 9443, PCC 9701, PCC 9717, PCC 9432, PCC 9809, PCC 9806 e SPC777. Foi utilizado um padrão de cores específico para a designação do continente de origem de cada linhagem, sendo ele: roxo = Ásia; rosa = Oceania; vermelho = África; amarelo = Europa; verde = América do Norte; e azul = América do Sul.

Fonte: Weiss (2017)

Desta forma, é evidente que as linhagens não apresentam quantidades e categorias iguais de agrupamentos relacionados a síntese de metabólitos secundários, ainda que similares. Contudo, devido a grande variação destas quantidades e categorias, não está claro se esta heterogeneidade é resultado de processos biológicos, ou mesmo se é um artefato resultado dos métodos genômicos.

6.3. Comparações genômicas globais

Para tentar elucidar se a heterogeneidade de quantidades e categorias de agrupamentos relacionados ao metabolismo secundário encontrada trata-se de processos biológicos ou metodológicos, análises globais comparativas dos genomas foram realizados. Tais análises compreendem estudos de sintenia e homologia genômicas, e análises quantitativas do pangenoma e do genoma cerne (coregenome).

6.3.1. Colinearidade entre as linhagens da espécie *Microcystis aeruginosa*#

A organização relativa, avaliada pela presença de blocos sintênicos colineares, mostrou-se bastante complexa (Figura 7), dificultando a observação de padrões. Nesta espécie, os blocos colineares são regiões pequenas, mostrando um padrão de intenso rearranjo cromossômico. Esta ordem complexa do cromossomo - ou a ausência de sintenia clara - da espécie pode ser reflexo de eventos de inserção frequentes, como documentados na literatura (LIN et al., 2011) e reaccessados neste trabalho.

De maneira generalizada, os cromossomos de espécies procariontes se mostram aptos a sofrer reorganizações globais com frequência. DARLING, MIKLÓS e RAGAN (2008), mostraram que, em *Yersinia pestis*, o rearranjo não somente é frequente, como também resulta em diferenças fenotípicas importantes, implicando em processos evolutivos que agem sobre tais rearranjos, como em mutações. Desta forma, esta é potencialmente uma das razões para o padrão aparentemente complexo de sintenia entre as sequências dos genomas das linhagens em estudo, e pode refletir em padrões genômicos de agrupamentos relacionados ao metabolismo secundário.

Por outro lado, é plausível que tal complexidade seja fruto das metodologias de sequenciamento e, em especial, montagem dos genomas, tendo em vista que apenas três dentre os dezoito genomas analisados apresentam-se completos e finalizados. Para averiguar esta suposição, um teste onde foi avaliada a sintenia somente entre duas linhagens de genomas completos (somente uma sequência contígua) foi realizado. Neste teste ainda pôde-se observar uma relação altamente intrincada de colinearidade, similar à análise considerando todas as 18 linhagens (Figura 8). Isso evidencia a sugestão de que tal padrão pode não ser produto de manipulações metodológicas, mas sim característica natural da espécie.

Ainda, o potencial de interferência nas análises comparativas proporcionado pelas diferentes quantidades de sequências contíguas de cada genoma foi averiguado estatisticamente, mais adiante neste trabalho.

Figura 8: Alinhamento para a detecção de blocos colineares sintênicos entre somente duas sequências completas (NIES-2549 e NIES-843). O padrão complexo ainda pode ser observado, o que aponta para o fato deste padrão ser natural da espécie, e não fruto de interferência dos métodos genômicos

Fonte: Weiss (2017)

6.3.2. Homologia de sequência genômica entre as linhagens da espécie *M. aeruginosa*#

Uma análise foi realizada comparando-se os 18 genomas, e utilizando como referência o genoma da linhagem NIES-843. A escolha da referência nesta análise foi baseada no fato de que esta linhagem é a mais estudada entre as 18 utilizadas no presente trabalho, pois é a primeira linhagem de *Microcystis aeruginosa* a ter sua sequência genômica completa publicada. Desta forma, esta comparação nos permite a apreciação de características da espécie reconhecidas na linhagem NIES-843, potencialmente refletidas na sequência genômica (Figura 9).

Fonte: Weiss (2017)

Figura 9: Mapa circular genômico da espécie *Microcystis aeruginosa*, comparando-se todas as linhagens com a linhagem NIES-843. Foi utilizado um padrão de cores específico para a designação do continente de origem de cada linhagem, de forma a permitir a avaliação de padrões biogeográficos de homologia entre as linhagens, caso algum: roxo = Ásia; rosa = Oceania; vermelho = África; amarelo = Europa; verde = América do Norte; e azul = América do Sul.

O mapa circular genômico evidencia muitas regiões bastante conservadas (apresentando entre 70% e 100% de identidade), as quais se mostram com cores mais intensas, entremeadas por muitas regiões relativamente pequenas, pouco conservadas ou demasiadamente dissimilares em relação à referência (abaixo de 50% de identidade, branco), apresentando conformidade com os resultados obtidos do estudo sintênico realizado com o *software* Mauve. Este padrão é, também, indicativo de alguma dinâmica genômica.

O mapa circular torna evidente a ausência deste padrão entre todos os continentes em relação à referência. Isso pode ser produto de uma participação pequena de cada continente na

amostragem, de maneira que a variabilidade e a comunalidade da espécie seriam pouco representadas pelos 18 genomas. Em contrapartida, pode ser o simples reflexo da ausência de variação biogeográfica nesta espécie – suposição substancial, pois é muito frequente entre micro-organismos.

GREMBERGHE e colegas (2011) mostram a ausência de padrões biogeográficos para a espécie *Microcystis aeruginosa*, porém valendo-se de sequências parciais (regiões ITS). Naquele estudo, os resultados mostram a ausência de estrutura filogeográfica entre as 311 sequências de rRNA-ITS dos 6 continentes, e ainda mostram uma relação filogenética marginal entre as sequências de mesmo continente, evidenciando que a espécie é realmente cosmopolita, mesmo que apresente padrão menos homogêneo de distribuição geográfica a nível intracontinental. O estudo de GREMBERGHE et al. (2011) corrobora os resultados aqui obtidos, pois sugere que a espécie *M. aeruginosa* apresenta um padrão genômico conservado entre continentes, mesmo que haja claras variações genômicas locais.

6.3.3. Estimativa quantitativa do pangenoma e do genoma cerne da espécie *M. aeruginosa*

A análise do pangenoma das linhagens estima que a espécie apresenta mais de 9.400 genes diferentes (Figura 10). HUMBERT e colegas (2013) mostraram uma estimativa de ao menos 12.000 genes diferentes para a espécie *Microcystis aeruginosa*, contudo seu estudo avalia somente 12 genomas de linhagens diferentes. A discrepância entre estes resultados evidencia a característica desta espécie que apresenta um genoma considerado “aberto” - que seria aquele em que ocorre observamos a incompletude das informações a respeito de todas os genes ou variações nas sequências gênicas e genômicas.

A linhagem PCC7005 não pôde ser utilizada para esta análise, considerando o funcionamento do algoritmo, que se baseia em procura por similaridade de sequência. Desta forma, possivelmente o grande número de sequências contíguas desta linhagem, que resulta em uma grande quantidade de genes fragmentados e leva à superestimativa dos valores, interferindo nas análises, não permite que suas sequências gênicas sejam apropriadamente avaliadas.

A representação gerou um padrão gráfico frequentemente encontrado em espécies cuja amostragem gênica é pequena em relação à totalidade. A curva para os 17 genomas não alcança a horizontalidade, nos levando a concluir que a amostragem não é suficiente para a estimar a quantidade total de genes da espécie, e que, portanto, a quantidade estimada de 9.400 genes é somente uma sugestão que ainda poderá ser melhor especulada com o

acrécimo de genomas considerados neste tipo de análise. Este resultado, porém, pode ser resultado de transferências horizontais frequentes e rearranjos.

O genoma cerne da espécie foi estimado em 1.944 genes a partir das 17 linhagens (Figura 11). O resultado mostra um grande número de genes se compararmos com a proporção de genes totais entre os genes do genoma cerne de algumas das linhagens - abrangendo entre 50% e 35%, aproximadamente. Contudo, a curva ajustada mostra uma horizontalidade constante, indicativa de uma estimativa próxima do real para a espécie, o que significa que há baixa probabilidade de esse número mudar significativamente com o acréscimo de genomas na amostragem.

Figura 10: Estimativa quantitativa do pangenoma da espécie *Microcystis aeruginosa* a partir de 17 genomas disponíveis para análise. A abcissa indica a quantidade de genomas reamostrados para a estimativa, e a ordenada indica o número estimado de genes. A curva ajustada mostra um pangenoma “aberto”, em que pode-se esperar que mais genomas aumentem o número de genes encontrados na espécie. Isso indica a grande diversidade gênica da espécie, ou (não excludentemente) que a amostragem é pequena para estimar o total de genes da espécie.

Fonte: Weiss (2017)

Considerando estes resultados, podemos concluir que os genomas utilizados nesta análise permitem a representação robusta e suficiente dos genes comuns à espécie *Microcystis aeruginosa*. Desta forma, fica evidente que novos sequenciamentos de linhagens desta espécie são desnecessários para a avaliação do conteúdo gênico mínimo definidor da espécie, ainda

que para a totalidade de genes mais sequenciamentos devem ser incluídos nas análises. Este fato é particularmente interessante para estudos de procura por agrupamentos gênicos desconhecidos.

Figura 11: Estimativa de genome cerne da espécie *Microcystis aeruginosa* a partir de 17 genomas. A abscissa mostra a quantidade de genes reamostrados para a estimativa, enquanto a ordenada mostra a quantidade estimada de genes. A assíntota estimada é de 1.944 genes. A horizontalidade evidente a partir do sétimo genome mostra robustez estatística na estimativa.
Fonte: Weiss (2017)

6.4. Descrição estatística, representatividade dos genomas, e inferências

Com o objetivo de verificar a normalidade da distribuição das observações amostrais, o teste Shapiro-Wilk foi utilizado, permitindo a classificação da distribuição das seis variáveis utilizadas neste trabalho. Em um primeiro momento, foram considerados os 18 genomas, e dessa maneira, cinco das seis variáveis puderam ser classificadas como distribuídas normalmente (Tabela 5). A variável quantidade de sequências contíguas foi a única a não se apresentar distribuída normalmente.

Tabela 5 – Teste de normalidade de Shapiro-Wilk aplicado às variáveis, tanto considerando a linhagem PCC 7005, quanto excluindo-a. Verificamos sua influência na distribuição estatística das outras variáveis, de maneira particular a cada uma. I: quantidade de transposases preditas; II: tamanho do genoma em megapares de bases; III: quantidade de genes preditos; IV: quantidade de agrupamentos relacionados a síntese de metabólitos secundários; V: quantidade de sequências contíguas; VI: quantidade de categorias de metabólitos secundários.

Variáveis	Teste (18 genomas)	Classificação	Teste (PCC 7005 excluída)	Classificação
I	0,030	Normal	0,024	Normal
II	0,676	Normal	0,698	Normal
III	0,537	Normal	0,574	Normal
IV	0,545	Normal	0,542	Normal
V	0,009	Não Normal	0,037	Normal
VI	0,134	Normal	0,014	Normal

Analisando pormenorizadamente as observações das amostras (Tabela 6), podemos verificar que a variável quantidade de sequências contíguas apresenta uma discrepância substancial na linhagem PCC 7005 (1.363 sequências contíguas, mais de 460% de afastamento da média). Isso sugere uma distorção na distribuição da variável por esta observação, o que acarretaria em distorções em todas as análises descritivas e inferenciais realizadas baseadas ou correlacionadas a esta variável, e implica em distorções potenciais em outras observações desta linhagem. Desta maneira, para as análises estatísticas realizadas, duas formas serão apresentadas: aquela que inclui a PCC 7005, e aquela que exclui a PCC 7005.

Tabela 6 – Observações de cada variável utilizadas nas análises estatísticas, para a caracterização quantitativa genômica da espécie *M. aeruginosa*. I: quantidade de transposases preditas; II: tamanho do genoma em megapares de bases; III: quantidade de genes preditos; IV: quantidade de agrupamentos relacionados a síntese de metabólitos secundários; V: quantidade de sequências contíguas; VI: quantidade de categorias de metabólitos secundários.

Linagens	I	II	III	IV	V	VI
SPC777	115	5,45	5194	12	278	7
PCC7005	na	4,90	4603	21	1.363	10
PCC9809	75	5,01	4695	18	303	7
PCC9806	65	4,26	3964	6	93	4
PCC9432	57	4,99	4719	15	132	8
PCC7941	49	4,80	4468	11	77	4
PCC9717	75	5,30	4922	17	264	8
PCC9701	49	4,75	4458	14	323	8
PCC9443	56	5,18	4863	16	221	8
PCC9807	60	5,15	4789	14	267	7
PCC9808	69	5,05	4808	14	141	7
DIANCHI905	118	4,85	4587	13	335	5
TAIHU98	361	4,84	4586	11	4	4
NIES-88	256	5,26	4969	11	262	6
NIES-44	37	4,56	4246	11	79	5
NIES-2481	35	4,29	3952	10	1	7
NIES-2549	102	4,29	3938	10	1	7
NIES-843	559	5,84	5484	11	1	6

A forma que exclui a linhagem PCC 7005 tem como objetivo a melhor representatividade das características naturais da espécie, descrevendo-a mais proximamente da realidade. A forma em que a linhagem PCC 7005 permanece, nos permite a comparação e observação da influência da quantidade de sequências contíguas pode exercer sobre as outras variáveis e descrições estatísticas, em especial sobre as características genômicas relacionadas ao metabolismo secundário.

Realizando um novo teste de normalidade em que excluiu-se a linhagem PCC 7005, podemos observar, primeiramente, que a variável quantidade de sequências contíguas apresenta-se distribuída normalmente, permitindo realizar as análises descritivas e inferenciais sem maiores distorções. Também, observamos que a simples exclusão desta linhagem causa

modificações diferentes em muitas das outras variáveis, o que em si aponta para a associação quantitativa da variável quantidade de sequências contíguas e as outras variáveis.

A exclusão mostra também que a quantidade de sequências contíguas possui relações estatísticas diferentes com as outras variáveis, desde muito evidente (como em relação à variável quantidade de categorias de metabólitos secundários), até muito sutil (como no caso da variável quantidade de agrupamentos gênicos) (Tabela 6). Desta maneira, análises de correlação são interessantes para o melhor descrição deste fenômeno.

6.4.1. Descrição estatística amostral, estimativas da espécie e correlações

A partir das variáveis utilizadas, podemos descrever quantitativamente a espécie *Microcystis aeruginosa* a partir da descrição da amostra composta pelos genomas em estudo. A descrição apresentada na Tabela 7 fornece um panorama das características genômicas da espécie e intervalo de confiança de 95%, inferidos a partir da amostragem.

Tabela 7 – Descrição quantitativa das variáveis analisadas. Intervalo de confiança considerado de 95%. Média amostral = μ . I: quantidade de transposases preditas; II: tamanho do genoma em megapares de bases; III: quantidade de genes preditos; IV: quantidade de agrupamentos relacionados a síntese de metabólitos secundários; V: quantidade de sequências contíguas; VI: quantidade de categorias de metabólitos secundários.

Variável	Média (18 genomas)	Desvio padrão (18 genomas)	Intervalo de confiança (18 genomas)	Média (PCC 7005 excluída)	Desvio padrão (PCC 7005 excluída)	Intervalo de confiança (PCC 7005 excluída)
I	118,77	139,23	61,87 - 175,68	125,76	140,23	66,58 - 184,94
II	4,93	0,41	4,76 - 5,10	4,93	0,43	4,75 - 5,11
III	4624,72	417,30	4454,16 - 4795,27	4626,00	430,10	4444,48 - 4807,51
IV	13,05	3,50	11,62 - 14,48	12,59	2,98	11,33 - 13,84
V	230,27	307,31	104,67 - 355,88	163,64	124,24	111,21 - 216,08
VI	6,55	1,65	5,88 - 7,23	6,35	1,45	5,73 - 6,96

A quantidade de transposases preditas para a espécie ($125,76 \pm 140,23$; intervalo de confiança de 95%) sugere uma dinâmica genômica intensa, potencialmente representando uma situação específica da espécie como um todo. Contudo, seu desvio padrão é substancialmente elevado, sendo mesmo maior que a média. Isso mostra uma larga amplitude de quantidades na amostra. De fato, as duas linhagens com maiores quantidades de transposases preditas são a NIES-843 (N = 559) e TAIHU-98 (N = 361). Desta maneira, é

plausível considerar a amostragem dos 18 genomas como pequena para avaliar esta característica da espécie ou população.

Ainda, devido à diferença evidentemente substancial destas linhagens na amostra, propõe-se que a quantidade de transposases preditas nas linhagens que se apresentam fora do intervalo de confiança (e, portanto, que se mostram extremáticas), TAIHU-98 (N = 361), NIES-843 (N = 559), NIES-2481 (N = 35), NIES-88 (N = 256), NIES-44 (N = 37), PCC 7941 (N = 49) e PCC 9701 (N = 49) seja resultado de processos particulares das linhagens, seja esta característica natural das linhagens, ou seja ela resultado de artefatos de técnicas microbiológicas ou genômicas.

As linhagens NIES-843, NIES-2481, NIES-88 e NIES-44, coletadas no mesmo ambiente – o Lago Kasumigaura, Japão – apresentam-se todas fora do intervalo de confiança da estimativa, além do padrão bastante marcado por extremos – duas muito acima da média amostral, e duas muito abaixo, e uma linhagem (NIES-2549) apresentando-se muito próxima à média amostral (N = 102). Isso sugere fortemente que o número de transposases, que varia significativamente entre os genomas sequenciados desta espécie, é característica específica de cada linhagem, apontando para processos de reorganização genômica e dinamismo diferentes e particulares em linhagens da mesma espécie, que ocorrem no mesmo ambiente.

SABART e colegas (2009) mostram diferenças genéticas entre linhagens parcialmente sequenciadas de duas represas interligadas por um canal, na França. As populações analisadas naquele trabalho apontam para uma estrutura complexa, em que um genótipo dominante entre todas as populações da espécie *Microcystis aeruginosa* evidencia a ausência de padrão biogeográfico, ainda que a diferença entre algumas das populações indica uma interrupção de fluxo gênico em uma das reservas, mesmo que os fluxos hídricos interconectem entre todas as áreas avaliadas, resultando em subpopulações dentro das áreas avaliadas. Os resultados de SABART et al. (2009) suportam a sugestão a respeito da complexidade populacional na medida em que também evidencia a tendência desta espécie em sofrer modificações locais, ainda mantendo uma estrutura genômica populacional contínua.

Interessantemente, as linhagens NIES-843 (N = 559, a maior quantidade de transposases preditas), NIES-2549 (muito próximo da média amostral) e NIES-2481 (N = 35, a menor quantidade de transposases preditas) são as três linhagens que atualmente apresentam sequenciamento completo na espécie *Microcystis aeruginosa*. Isto reforça a proposição que se sugere que esta diferença não é produto de artefatos metodológicos de sequenciamento e montagem, mas sim esta diferença é resultado de processos biológicos.

Dessa forma, fica claro que as estimativas de quantidade de transposases preditas para a população a partir da amostra não é representativa da espécie, mas sugere uma dinâmica de intensidade bastante variada de inserções (e potencialmente de rearranjos como resultado) como possível característica da espécie (HUMBERT et al., 2013). Além disso, sugere-se a influência do componente microecológico na organização genômica específica a cada linhagem para a espécie *Microcystis aeruginosa*.

A média amostral da variável tamanho de cada genoma ($4,93 \pm 0,43$; intervalo de confiança de 95%) apresenta-se como um tamanho considerado mediano para cianobactérias. Também, seu coeficiente de variação (8,44% – substancialmente baixo) é o menor entre as seis variáveis avaliadas, o que mostra a relativa homogeneidade nos tamanhos dos genomas da amostra. E, de fato, os tamanhos de todos os genomas são abarcados no intervalo de confiança, mostrando claramente que esta variável da amostra é capaz de representar substancialmente a espécie *Microcystis aeruginosa*, sendo a estimativa da média do tamanho do genoma para a espécie bastante próxima da média da amostra.

O baixo coeficiente de variação também é indicativo de que o tamanho do genoma das diferentes linhagens de *Microcystis aeruginosa* mantém-se constante, mesmo frente a diferentes pressões ecológico-evolutivas sugeridas pela variável quantidade de transposases. A associação entre o tamanho do genoma e a quantidade de transposases (mais à frente) pode oferecer uma sugestão de cenário evolutivo para a espécie.

A variável quantidade de genes preditos para cada genoma apresenta média $4626 \pm 430,10$ (intervalo de confiança de 95%), e coeficiente de variação de 9,02%, mostrando relativa homogeneidade da quantidade de genes na amostra. Considerando que todas as linhagens estão abarcadas dentro do intervalo de confiança desta variável, podemos aceitar esta estimativa de média como representante da espécie, mesmo para uma amostragem considerada pequena na perspectiva estatística.

Como esperado, é observável coincidências claras entre a quantidade de genes e tamanho do genoma. A exemplo, as linhagens de maior (NIES-843, N = 5484) e de menor (NIES-2549, N = 3938) quantidade de genes preditos são igualmente aquelas de maior e menor tamanho de genoma. Além disso, ambas as variáveis apresentam coeficiente de variação muito similares. Isso evidencia fortemente, como esperado, a associação entre o tamanho do genoma e a quantidade de genes preditos. Essa associação também foi testada através de análises de correlação linear, mais adiante.

A variável quantidade de agrupamentos gênicos relacionados a metabólitos secundários preditos para cada genoma mostra uma média de $12,58 \pm 2,98$ (intervalo de

confiança de 95%) agrupamentos gênicos para cada genoma. Seu coeficiente de variação é 26,85%, apresentando portanto heterogeneidade, ou relativa pouca regularidade na quantidade de agrupamentos.

A variável quantidade de sequências contíguas definidas no sequenciamento disponível para cada genoma apresenta uma média amostral de $163,64 \pm 124,24$ (intervalo de confiança de 95%) sequências contíguas por genoma. Considerando que entre os genomas utilizados para estas estimativas, há três que apresentam sequência completa, consistindo, portanto, como uma única sequência contígua (NIES-843, NIES-2549 e NIES-2481), o desvio padrão elevado é esperado. Ainda, a linhagem TAIHU98 é caracterizada por somente quatro sequências contíguas.

A linhagem PCC 7005 é caracterizada por uma quantidade muito grande de sequências contíguas; muito acima da média amostral ($n = 1363$; afastamento da média de 466,86%), e claramente causa uma deformação na distribuição normal esperada para a variável, como anteriormente tratado. Esta linhagem também mostra desvios em outra variável analisada (quantidade de agrupamentos gênicos relacionados ao metabolismo secundário), além da impossibilidade de análise a respeito de quantidade de transposases preditas devido, possivelmente, ao grande número de sequências contíguas (o que dificulta a predição por homologia de genes), e ainda, potencialmente pelas mesmas razões, a impossibilidade de análise de pangenoma e genoma cerne, apresentados anteriormente.

A possível causa da relação entre os artefatos gerados nas outras variáveis e a quantidade de sequências contíguas é o fato de que maior quantidade de sequências contíguas é sinônimo de fragmentos de sequência genômica mais curtos, levando ao aumento da quantidade de genes artificialmente interrompidos, incompletos, ausentes, ou inaptos para a identificação. De maneira inversa, quanto menor a quantidade de sequências contíguas, menor é a quantidade de genes interrompidos, melhorando a contagem total de genes, a identificação destes, e a representatividade da sequência genômica. Regiões repetitivas do genoma são frequentemente as causas para dificuldades no processo de montagem de sequências genômicas, e estas regiões são frequentes nos genomas cianobacterianos, o que explica (ao menos em parte) a dificuldade especificamente inerente de sequenciamento e montagem dos genomas cianobacterianos.

Fica claro, ainda, baseado nesta comparação de amostragens no teste de normalidade e pela observação caso a caso dos dados, que a quantidade de sequências contíguas tem o potencial de interferir na análise de algumas variáveis, quando referem-se a estruturas bastante específicas, como é o caso dos agrupamentos gênicos. Nesses casos, a sequência e o

ordenamento dos genes são essenciais para o entendimento de sua evolução, atividades, funções e identificação. Desta maneira, uma simulação da quantidade de sequências contíguas e a quantidade de agrupamentos localizados foi realizada, e é apresentada mais à frente.

Contudo, sequências contíguas na casa das poucas centenas ainda podem oferecer um panorama pouco enviesado pelas metodologias, sendo usualmente suficiente para o estudo de estruturas ou processos que independem de organizações de relação não biunívoca, como quantidade de genes preditos e tamanho dos genomas.

A variável quantidade de categorias de metabólitos preditos genomicamente mostra uma média de $6,35 \pm 1,45$ (intervalo de confiança de 95%) categorias de metabólitos preditos por genoma. Seu coeficiente de variação é de 12,21%, sendo, portanto, uma variação bastante reduzida em relação a outras variáveis consideradas neste estudo. Isso mostra que a amostragem é representativa da quantidade de categorias que poderíamos encontrar na espécie.

Contudo, é interessante notar que a variação em quantidade de categorias observado na espécie *Microcystis aeruginosa* é peculiar, e, portanto, as razões para essa variação, mesmo que baixa em relação às outras variáveis, é objeto interessante para aprofundamento nas investigações, inclusive sendo descritas através de análises como associação desta variável em relação a outras, como quantidade de sequências contíguas ou quantidade de agrupamentos relacionados ao metabolismo secundário. Também é interessante notar que a classificação de agrupamentos em categorias de metabólitos a que referem-se é baseada, pela plataforma antiSMASH, na assinatura específica de cada categoria. Isso quer dizer que um padrão gênico encontrado em um agrupamento deve obedecer estritamente ao padrão definido de homologia e ordenamento dos genes contidos no agrupamento referência. Este tipo de análise promove a confiabilidade de previsões e identificações, mas também gera algumas dificuldades, como algumas estimativas acima ou abaixo do real.

Fica claro que, em genomas com muito alto número de sequências contíguas, como é o caso da PCC 7005, o número de categorias é bastante elevado, mas isso não deve refletir a natureza real da linhagem. É consistente supor que se trate de um resultado de viés, produto da partição múltipla de agrupamentos originais em agrupamentos menores e identificados automaticamente como outras categorias, alterando artificialmente o número de categorias (além de outras variáveis), sem apresentar, contudo, significado biológico factual.

Em contrapartida, a linhagem TAIHU98, que apresenta somente quatro sequências contíguas, mostrou um número substancialmente baixo ($N = 4$) de categorias, enquanto nas linhagens como NIES-843 e NIES-2549, que apresentam o genoma sequenciado

completamente, e portanto não apresentam vieses associados à quantidade de sequências contíguas, foram encontrados 6 e 7 categorias de metabólitos, respectivamente. Ainda que a diferença entre estas linhagens (e entre as outras, de forma geral) não seja elevada, ela ainda é presente e parece ser uma característica da espécie. Isso evidencia a variabilidade de quantidades de categorias de metabólitos que podem ser encontradas na espécie *Microcystis aeruginosa*.

6.4.2. Correlações das estatísticas descritivas amostrais

A partir das variáveis analisadas, correlações lineares podem ser delineadas afim de evidenciar associações quantitativas. Em alguns pontos durante a descrição das variáveis, foi proposto que correlações podem indicar associações que apontem a respostas a algumas hipóteses. Desta maneira, comparamos também os conjuntos amostrais que incluem e que excluem a linhagem PCC 7005, testando o efeito de variáveis que não seguem a distribuição normal sobre análises descritivas e inferenciais. A Tabela 8 apresenta os resultados, e a Tabela 9 apresenta as categorias utilizadas para classificação das correlações.

Esta seção considera as correlações entre as variáveis aleatórias observadas na amostra, e propõe explicações nos casos de correlações mais fortes e moderadas. Admitidamente, as associações estatísticas não comprovam nenhuma hipótese, somente corroboram ou limitam explicações de padrões encontrados nas análises, e convidam a aprofundamentos e experimentações.

Tabela 8 – Correlações lineares de Pearson (R) entre as seis variáveis avaliada.

Variáveis correlacionadas	Correlação R (18 genomas)	Classificação (18 genomas)	Correlação R (PCC 7005 excluída)	Classificação (PCC 7005 excluída)
Quant. transp. x Tam. gen.	+ 0,527	moderada	+ 0,535	moderada
Quant. transp. x Quant. genes	+ 0,524	moderada	+ 0,534	moderada
Quant. transp. x Quant. agru.	- 0,299	fraca	- 0,222	fraca
Quant.transp. x Quant. seq.con.	- 0,322	fraca	- 0,331	fraca
Quant.transp. x Quant.cat.met.	- 0,348	fraca	- 0,284	fraca
Tam.gen x Quant. genes	+ 0,995	muito forte	+ 0,995	muito forte
Tam.gen. x Quant. Agrup.	+ 0,377	fraca	+ 0,471	moderada
Tam.gen. x Quant. seq. con.	+ 0,118	muito fraca	+ 0,345	Fraca
Tam.gen. x Quant. cat. met.	+ 0,235	fraca	+ 0,287	fraca
Quant. genes x Quant. agru.	+ 0,374	fraca	+ 0,463	moderada
Quant.genes x Quant.seq. com.	+ 0,129	muito fraca	+ 0,361	fraca
Quant. genes. x Quant.cat.met.	+ 0,221	fraca	+ 0,267	fraca
Quant.agru. x Quant.seq.con.	+ 0,716	forte	+ 0,606	Moderada
Quant. agru. x Quant.cat.met.	+ 0,765	forte	+ 0,669	moderada
Quant.seq.con. x Quant.cat.met.	+ 0,608	moderada	+ 0,389	fraca

Tabela 9 – Intervalos da classificação utilizada para as correlações lineares.

Amplitude	Classificação
0,00 – 0,19	Muito fraca
0,20 – 0,39	Moderadamente fraca
0,40 – 0,69	Moderada
0,70 – 0,89	Moderadamente forte
0,90 – 1,00	Muito forte

A comparação dos resultados obtidos para as correlações entre conjuntos amostrais que inclui e que exclui a linhagem PCC 7005, podemos verificar que houve uma redução de classificações extremas. Isso evidencia a importância da verificação da distribuição normal para as variáveis que serão testadas para correlação linear. Ainda, algumas correlações mais intermediárias ou fracas tornaram-se mais fortes, também refletindo a importância do teste Shapiro-Wilk de normalidade para correlações mais moderadas.

No caso das correlações agrupamentos gênicos relacionados ao metabolismo secundário x quantidade de sequências contíguas, e quantidade de categorias de metabólitos

secundários x quantidade de sequências contíguas (ambas fortes para o conjunto que inclui a linhagem PCC 7005, porém moderadas para o conjunto que exclui a linhagem PCC7005), a análise é de fundamental importância para este trabalho, e sem a verificação de aderência à distribuição normal, os valores podem ser obtidos com menor confiabilidade, e os resultados podem ser mal interpretados.

De mesma maneira, as correlações tamanho de cada genoma em Megapares de bases x quantidade de agrupamentos gênicos relacionados ao metabolismo secundário e quantidade de genes preditos x quantidade de agrupamentos gênicos relacionado ao metabolismo secundário, são fracas no conjunto que inclui a linhagem PCC 7005, porém moderadas para o conjunto que exclui a linhagem PCC 7005, assim tornando-se mais relevantes para o estudo desta espécie. Portanto, em uma situação em que uma variável não se mostra devidamente amostrada, temos a interpretação equivocada de ausência de associação entre estas variáveis, contudo em realidade há associação estatística, mesmo que somente moderada.

Apesar de correlações lineares não mostrarem, evidentemente, a mecânica ou causalidade explícita em um processo biológico, podem ainda indicar tendências, portanto colaboram na sugestão de hipóteses explicativas para padrões. Além disso, permitem a formulação de novas hipóteses que podem ser testadas experimentalmente, em especial quando as correlações são muito fortes (0,90 a 1,00) ou fortes (0,89 a 0,7) - quando somente entre 0% e 49% da variação de uma variável não pode ser explicada pela outra variável.

A lógica por trás das correlações moderadas a fortes pode ser definida por três possibilidades excludentes: i) as mudanças em uma variável implica na outra, reciprocamente, portanto as duas variáveis compartilham uma origem causal, mesmo que biologicamente parcial; ii) as mudanças em uma variável decorre ou é decorrente das mudanças em outra variável, o que implica em causalidade direta ou indireta entre as duas variáveis; e iii) A relação de proporcionalidade da correlação entre as duas variáveis é uma coincidência.

Nos casos de correlações moderadas (0,4 a 0,69), há indícios suficientes para a construção de hipóteses a respeito da potencial interação causal entre dois processos, contudo, de maneira geral, pode indicar a confluência de mais fatores do que aqueles que estão sendo levados em consideração na construção da correlação (neste intervalo, aproximadamente entre 50% e 84% da variação de uma variável não pode ser explicada pela outra variável), ou simplesmente indicando coincidências – portanto, limitando as inferências derivadas de tal análise.

Ainda, nas correlações fracas (0,20 a 0,39) e muito fracas (0,0 a 0,19), quando entre 85% e 100% da variação de uma variável não pode ser explicada pela outra variável, as

observações mostram a não interferência entre dois processos em seus resultados mensurados (as variáveis), portanto não permitindo a inferência de processos ou levantamento de hipóteses embasadas nas observações.

6.4.3. Aumento do genoma mediado por transposases

Considerando-se a correlação entre a quantidade de transposases preditas em cada genoma x tamanho de cada genoma em Megapares de bases, de intensidade moderada (+0,535), podemos verificar a associação entre o tamanho do genoma (em pares de base) e a quantidade de transposases preditas para cada genoma. Somando-se a isso, verificamos a grande variação da quantidade de transposases preditas para cada linhagem, permitindo propor que há aumento substancial de sequência genômica resultante de eventos de inserção. Contudo, esta associação mostra que a inserção mediada por transposases não é o único processo envolvido no aumento da sequência genômica. Além disso, o acréscimo pode ser constituído por sequências funcionais e não funcionais.

A correlação entre tamanho de cada genoma x quantidade de genes preditos em cada genoma, de intensidade muito forte (+0,995), indica que o tamanho do genoma nas linhagens de *Microcystis aeruginosa* é acompanhado muito proximamente da quantidade de genes. Isso evidencia estatisticamente o acúmulo de sequências funcionais pelos eventos de inserção proporcionados pela atividade das transposases, além de outros processos não avaliados que aumentam o genoma. Estas sequências inseridas, contudo, podem ser tanto da própria transposase (pois são categorizados como funcionais), quanto de genes vizinhos às transposases, os quais podem ser duplicados e inseridos juntamente com o gene da transposase.

A correlação entre a quantidade de transposases preditas x quantidade de genes preditos, que associa a quantidade de transposases e quantidade de genes, de intensidade moderada (+0,533), mostra que o aumento de sequência, e portanto a inserção de genes realizada pela atividade das transposases, pode ser causada tanto pela inserção de sequências da própria transposase, como também por sequências de outros genes. Isso sugere que a inserção de genes pela atividade da transposase é um processo relevante para a manutenção do tamanho do genoma da espécie *Microcystis aeruginosa*, potencialmente com capacidade em aumentar a quantidade de sequência funcional do genoma.

É interessante notar que tanto o tamanho do genoma quanto a quantidade de genes preditos estão moderadamente correlacionados (+0,463 e +0,471, respectivamente) com a quantidade de agrupamentos gênicos relacionados ao metabolismo secundário, somando-se o

pangenoma “aberto”, sugerem que um genoma maior é indicativo de maior quantidade de agrupamentos, fato importante para pesquisas futuras relacionadas à procura de agrupamentos em genomas da espécie *Microcystis aeruginosa*. Contudo, as associações entre categorias de metabólitos, e tamanhos dos genomas e genes preditos, se mostram fracas, sugerindo que genomas maiores não devem conter mais (nem menos) categorias de agrupamentos gênicos.

6.4.4. Interferência dos métodos genômicos nos estudos comparativos

A interferência proporcionada pelos métodos de sequenciamento e montagem foram verificados através de associações entre a variável quantidade de sequências contíguas, e as outras variáveis. Para isso, pressupomos que a quantidade de sequências contíguas é indicador inverso da representatividade do sequenciamento em relação ao genoma – quanto menor a quantidade de sequências contíguas, mais próximo da sequência real.

A correlação entre a quantidade de agrupamentos gênicos relacionados ao metabolismo secundário x quantidade de sequências contíguas, de intensidade moderada (+0,606), indica a interferência dos métodos genômicos na avaliação dos agrupamentos, mas deixa claro que há outros fatores desconhecidos que interferem nesta associação. Isso é possivelmente resultado de processos como i) a interrupção da sequência do agrupamento (que usualmente ultrapassa os 30 kpb, portanto, a chance de ocorrer interrupção na montagem é substancial); ii) em contrapartida, a raridade relativa dos agrupamentos na sequência genômica reduz substancialmente a probabilidade de ocorrer interrupção do agrupamento em genomas com menores quantidades de sequências contíguas; e iii) do método de procura da plataforma antiSMASH, que depende da procura de similaridade de sequência e distância entre os genes que compõem o agrupamento. Assim, no evento de uma interrupção de um agrupamento gênico resultante de uma montagem, a quantidade de agrupamentos é superestimada, potencialmente de mesma maneira com a quantidade de categorias de metabólitos.

Apesar dessas associações mostrarem interferência dos métodos nas análises comparativas, é interessante notar que em outras correlações é evidente a ausência da interferência dos métodos nas comparações.

A correlação entre a quantidade de sequências contíguas x quantidade de categorias de metabólitos (+0,389) mostra que, mesmo que a interrupção de agrupamentos ocorra, a quantidade de categorias não se altera frequentemente, indicando que o método da plataforma antiSMASH é eficiente para identificar, em média, as categorias dos metabólitos localizados

mesmo em circunstâncias em que o genoma se apresenta com grande quantidade de sequências contíguas.

Estes resultados deixam evidente que os procedimentos genômicos para sequenciamento e montagem de sequência podem interferir, em maior ou menor intensidade, nas comparações das diversas características genômicas de qualquer espécie ou táxon, dependendo das características em avaliação. Também, aponta para a importância da ponderação no momento da tomada de decisão sobre aplicação de recursos em procedimentos que melhorem a montagem de uma linhagem já sequenciada, ou se tais recursos seriam mais bem aplicados ao serem dirigidos para o sequenciamento de novas linhagens, ampliando a abrangência e representatividade das análises comparativas, notoriamente ao considerarmos que a melhor aplicação desses recursos pode colaborar substancialmente com a qualidade dos trabalhos computacionais e genômicos, permitindo a construção de hipóteses mais robustas ou significativas.

Considerando a importância de tal tomada de decisão, o entendimento do comportamento da contagem de agrupamentos gênicos localizados e a fragmentação do genoma é fundamental. Para isso, foi realizado um experimento *in silico* onde foi simulado a fragmentação dos três genomas completos utilizados neste estudo (NIES-843, NIES-2481 e NIES-2549). Por serem completos, assume-se que estes genomas representam fidedignamente a situação real dos genomas *in vivo* (Figura 12).

Figura 12: Quantidade de agrupamentos gênicos relacionados ao metabolismo secundário localizados por quantidade de fragmentos simulados in silico, para a espécie *Microcystis aeruginosa*. A – linhagem NIES-843; B – linhagem NIES-2481; C – linhagem NIES-2549. Os três genomas utilizados neste experimento possuem somente 1 sequência contígua originalmente. Observe que NIES-2481 e NIES-2549 possuem originalmente 10 agrupamentos cada, e na região de inflexão alcançam aproximadamente 21 agrupamentos (B e C), enquanto a linhagem NIES-843 possui originalmente 11 agrupamentos, mas na região de inflexão a contagem alcança mais de 30 agrupamentos (A).

Podemos verificar claramente na Figura 12 o comportamento não linear entre a quantidade de agrupamentos localizados e a quantidade de fragmentos simulados. Isso explica a não normalidade da variável quantidade de sequências contíguas quando levamos em consideração a linhagem PCC7005. A inflexão da relação dos fragmentos e da quantidade de agrupamentos localizados das três linhagens utilizadas na simulação (e potencialmente das outras linhagens) está entre 900 e 1.000 fragmentos, e a linhagem PCC7005 é a única na amostra em que a quantidade de sequências contíguas é superior a este valor (especificamente 1.363).

Desta maneira, a linhagem PCC7005 causa desvios na distribuição normal da quantidade de sequências contíguas, além de causar também deformações estatísticas nas variáveis que são moderadas ou fortemente correlacionadas com a variável quantidade de sequências contíguas. Isso demonstra quanto os métodos genômicos podem interferir nas análises, e em especial, o quanto desvios podem ser causados por montagens genômicas incompletas.

Fica claro, portanto, que a variação da quantidade de agrupamentos relacionados a síntese de metabólitos observada neste trabalho é substancialmente influenciada pelos métodos genômicos. Mais especificamente, a descontinuidade dos genomas em casos em que a montagem não permite a resolução mais fidedigna da informação e da organização genômicas, possui uma relação não linear com a quantidade de agrupamentos localizados e resulta em análises comparativas bastante dificultosas.

Métodos matemáticos mais complexos poderiam ser aplicados de maneira a corrigir estes artefatos, caso um padrão de desvio fosse observado. Contudo, como podemos observar pela Figura 12, o padrão do comportamento da fragmentação sobre a contagem de agrupamentos é complexo, podendo sofrer influência de vários fatores, como densidade gênica no cromossomo, posição física dos agrupamentos, estrutura dos agrupamentos, geometria da fragmentação, tamanho médio das sequências contíguas, e mesmo os métodos específicos do programa utilizado para a localização dos genes, entre possivelmente outros. Podemos verificar um padrão gráfico na Figura 12, contudo o acesso a este é claramente difícil.

Desta maneira, a verdadeira variação da quantidade de agrupamentos localizados nas linhagens da espécie *Microcystis aeruginosa* é produto, ao menos em parte, dos métodos utilizados para o estudo do genoma e dos agrupamentos. Apesar de algumas linhagens apresentarem variação real (por exemplo, linhagem TAIHU98 com 4 agrupamentos localizados e somente 4 sequências contíguas, e NIES-843 com somente 1 sequência contígua

e 11 agrupamentos localizados), a variação encontrada nas outras linhagens com muitas sequências contíguas não pode ser considerada como de origem exclusivamente biológica, pois sofrem influência da quantidade de sequências contíguas. Pela figura 12 podemos observar que acima de aproximadamente 100 sequências contíguas encontramos um desvio não linear substancial, e desta maneira ainda não podemos diferenciar claramente os fatores de influência biológica e metodológica – contudo podemos apontar para o fato da dificuldade de discernimento entre as duas influências.

7. CONCLUSÕES#

Os tipos de metabólitos mais frequentemente encontrados na espécie *Microcystis aeruginosa* são bacteriocinas, terpenos, moléculas híbridas e NRPS. Dentre os peptídeos não ribossômicos, a aeruginosina, a microcistina e a micropeptina foram as mais frequentemente encontradas, apesar de nenhuma ser comum a todas as linhagens da espécie. Ainda, há substancial variação na similaridade, com relação aos padrões, dentro das sequências destes agrupamentos. Em alguns casos, o agrupamento se mostra tão modificado que, pelos métodos genômicos somente, é impossível discernir se tais variações se tratam de perda de ordenamento e funcionalidade, ou novas variações de agrupamentos conhecidos, ou ainda agrupamentos desconhecidos. Um agrupamento classificado pela plataforma antiSMASH como similar à micropeptina, apesar de conter algumas das estruturas esperadas para a potencial síntese de uma molécula, apresenta fatores de confundimento, como identidades baixas e ordenamento incomum, sendo necessária a proposição de estudos que se valham de outras estratégia complementares, como estudos de procura de elicitores para o ativamento do potencial agrupamento, e outras estratégias microbiológicas, biomoleculares ou bioquímicas.

A espécie *Microcystis aeruginosa* mostra uma dinâmica genômica intensa. Essa dinâmica não somente é refletida pela sentença tênue entre linhagens, mas também pela presença altamente heterogênea de transposases, as quais mostram regiões de inserção em quantidades evidentemente diferentes entre linhagens. A presença de transposases está estatisticamente associada com aumento, tanto em pares de bases quanto em número de genes, do genoma na espécie, sugerindo causalidade de processos de inserção e acréscimo de sequência de significado biológico. Tal intensidade e variabilidade de rearranjo e de inserções podem ser causas da variedade de agrupamentos relacionados ao metabolismo secundário e às classes estruturais de seus produtos preditos.

Contudo, a avaliação da interferência das inserções e rearranjos potencialmente promovidos pela atividade das transposases é substancialmente prejudicada pela amplitude das quantidades de sequências contíguas encontradas nos sequenciamentos dos genomas de cada linhagem. Essa heterogeneidade de sequências contíguas também interfere na avaliação da quantidade de agrupamentos relacionados ao metabolismo secundário, evidenciando a importância da qualidade das montagens proporcionadas nos bancos de acesso públicos.

Essa interferência nos estudos de agrupamentos, apesar de substancial, oferece dificuldades para análises automatizadas de agrupamentos relacionados ao metabolismo secundário que se baseiam em genomas acima de aproximadamente 100 sequências

contíguas para os genomas de *M. aeruginosa*. Acima deste limite há a possibilidade de estudos, contudo procedimentos e metodologias apropriadas devem ser adotadas, de maneira a verificar pormenorizadamente as asserções automatizadas.

A relação não linear entre a quantidade de agrupamentos e a quantidade de sequências contíguas, testada através de uma experimento *in silico*, sugere que proposições sobre a variabilidade da quantidade de agrupamentos preditos sofrem deformações metodológicas substanciais que impossibilitam, ou tornam altamente dificultosas, as análises comparativas desta característica, e de outras relacionadas, em genomas incompletos. Também mostra que a relação possivelmente segue um padrão, contudo muitas das potenciais variáveis são de difícil análise.

A média de sequências contíguas encontradas nos genomas utilizados mostra-se pouco acima de 163, indicando, de maneira generalizada, que estudos de agrupamentos gênicos relacionados ao metabolismo secundário devem sofrer vieses em análises automatizadas (contudo, esta avaliação deva ser levada caso a caso). A quantidade de sequências contíguas, ainda, não mostra grande interferência em estudos a respeito do tamanho e da quantidade de genes preditos, assim como apresenta pouca associação com a quantidade de categorias de metabólitos secundários preditos para os agrupamentos gênicos.

A espécie apresenta um genoma aproximado de 4,93 Mpb, contendo em média 4626 genes preditos. A associação tamanho-informação para o genoma desta espécie é forte, mostrando economia de sequência. Juntamente com a observação da alta dinâmica de inserção e rearranjo, além da pequena variação do tamanho dos genomas, podemos propor que a espécie mostra uma estratégia evolutiva de manutenção do tamanho e da informação contida no genoma, enquanto há uma dependência ou tolerância aos processos reorganizadores do ordenamento dos genes, havendo variações mesmo em linhagens de mesma localidade geográfica. Em uma perspectiva teleológica, isso pode ser produto também dos desafios variados oferecidos por sua distribuição cosmopolita. O genoma cerne, cuja estimativa se mostra bastante robusta com o tamanho da amostragem aqui utilizada, é relativamente elevado, sendo igual a 1.944 genes (perfazendo de 35% a 50% dos genes preditos, dependendo da linhagem). Isso corrobora a sugestão de manutenção do tamanho-informação sob efeito de rearranjos e inserções.

Ainda que potencialmente sob influência das quantidades de sequências contíguas, as linhagens da espécie *Microcystis aeruginosa* apresentam uma média aproximada de 12,5 agrupamentos gênicos relacionados com a síntese de moléculas bioativas, divididas em 6 ou 7 tipos diferentes de metabólitos secundários. O pangenoma avaliado mostra uma configuração

classificada como “aberto”, mostrando que ainda são necessários mais sequências genômicas para avaliar a totalidade de genes da espécie. Isso é interessante para prospecção de agrupamentos relacionados ao metabolismo secundário, sobretudo considerando que a associação entre genes preditos e quantidade de agrupamentos se mostra moderada para a espécie *M. aeruginosa*.

REFERÊNCIAS#

ALIKHAN, N.; PETTY, N.K.; ZAKOUR, N.L.B.; BEATSON, S.A. **BLAST Ring Image Genertor (BRIG): simple prokariote genome comparisions**. BMC Genomics, Londres, 12(402): 1-10, 2011.

APELDOORN, M.E. VAN; EGMOND, H.P. VAN; SPEIJERS, G.J.A.; BAKKER, G.J.I. **Toxins of cyanobacteria**. Molecular Nutrition and Food Research, Weinheim, 51(1): 7-60, 2007.

BLAZEJ, P.; MACKIEWICZ, P.; WANCZYK, M.; CEBRAT, S. **Evolution of bacterial genome under changing mutation pressure – Computer simulation studies**. Proceedings of the BIOSTEC, Varsóvia, 272-278, 2013.

BÖRNER, T.; DITTMAN, E. **Molecular biology of cyanobacterial toxins**. In: Huisman, J.; Matthijs, H.C.P.; Visser, P.M. (Ed.). **Harmful cyanobacteria**. Dordrecht: Springer, 2005.

BROCKS, J.J.; LOGAN, G.A.; BUICK, R.; SUMMONS, R.E. **Archean molecular fossils and the early rise of eukaryotes**. Science, Washington, 285(543): 1033–1336, 1999.

BERMAN-FRANK, I.; LUNDGREN, P.; FALKOWSKI, P. **Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria**. Research in Microbiology, Paris, 154: 157-164, 2003.

BUICK, R. **The antiquity of oxygenic photosynthesis: evidence from stromatolites in sulphate-deficient Archean lakes**. Science, Washington, 255(5040): 74-77, 1992.

CADEL-SIX, S.; DAUGA, C.; CASTETS, A.M.; RIPPKA, R.; MARSAC, T.; WELKER, M. **Halogenase genes in two non-ribosomal peptide synthetase gene clusters of *Microcystis* (Cyanobacteria): sporadic distribution and evolution**. Molecular Biological Evolution, Oxford, 25: 2031-2041, 2008.

CANNEL, R.J.P. **Algae as a source of biologically active products.** Pesticide Science, Toquio, 39: 147-153, 1993.

CASTENHOLZ, R.W. **General characteristics of the cyanobacteria.** *In*: Boone, D.R.; Castenholz, R.W. (Ed.), **Bergey's manual of systematic bacteriology.** 2a ed. New York, 474-487, 2001.

CADEL-SIX, S.; DAUGA, C.; CASTETS, A.M.; RIPPKA, R.; BOUCHIER, C.; MARSAC, T.; WELKER, M. **Halogenase genes in nonribosomal peptide synthetase gene clusters of *Microcystis* (cyanobacteria): sporadic distribution and evolution .** Molecular Biological Evolution, Oxford, 25(9): 2031–2041, 2008.

CHORUS, I.; BARTRAM, J. **Toxic cyanobacteria in water, a guide to their public health Consequences, monitoring and management.** London, World Health Organization, 416 p, 1999.

CONTRERAS-MOREIRA, B.; VINUESA, P. **GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis.** Applied and Environmental Microbiology, Washington, 79(24): 7696-7701, 2013.

DARLING, A.C.E.; MAU, B.; BLATTNER, F.R.; PERNA, N.T. **Mauve: multiple alignment of conserved genomic sequence with rearrangements.** Genome Research, Cold Spring Harbor, 14: 1394-1403, 2004.

DARLING, A.E.; MIKLÓS, I.; RAGAN, M.A. **Dynamics of Genome Rearrangements in Bacterial Populations.** Plos Genetics, San Francisco, 4(7): 1-16, 2008.

DARLING, A.E.; MAU, B.; PERNA, N.T. **progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement.** Plos One, San Francisco, 5(6): 1-17, 2010.

DIAS, J.M.A.; GONZALEZ, R.; FERREIRA, O. **Natural versus anthropic causes in variations of sand export from rivers: an example from the Guadiana river mouth (Southwestern Iberia).** Polish Geological Institute Special Papers, Varsóvia, 11: 95-102, 2011.

DITTMAN, E.; NEILAN, B.A.; BÖRNER, T. **Molecular biology of the peptide and polyketide biosynthesis in bacteria.** Applied Microbiology and Biotechnology, London, 57: 467-473, 2001.

DITTMANN, E.; GUGGER, M.; SIVONEN, K.; FEWER, D.P. **Natural product biosynthetic diversity and comparative genomics of the cyanobacteria.** Trends in Microbiology, Maryland Heights, 23 (10): 642-652, 2015.

DONIA, M.S.; SCHMIDT, E.W. **Linking chemistry and genetics in the growing cyanobactin natural products family.** Chemistry & Biology, London, 18: 508-519, 2011.

DU, L.; SHEN, B. **Identification and characterization of a type II peptidyl carrier protein from the bleomycin producer *Streptomyces verticillus* ATCC 15003.** Chemistry & Biology, Londres, 6: 507-517, 1999.

FELNAGLE, E.A.; JACKSON, E.E.; CHAN, Y.A.; PODEVELS, A.M.; BERTI, A.D.; MCMAHON, M.D.; THOMAS, M.G. **Nonribosomal peptide synthetases involved in the production of medically relevant natural products.** Molecular Pharmacology, Rockville Pike, 5(2): 191–211, 2008.

FINKING, R.; MARAHIEL, M.A. **Biosynthesis of nonribosomal peptides.** Annual Reviews of Microbiology, Palo Alto, 58: 453-488, 2004.

FIORE, M.F.; HONDA, R.Y. **Fixação do N₂ por Cianobactérias.** In: Figueiredo, M.V.B.; Burity, H.A.; Stamford, N.P.; Santos, C.E.R.S. **Microrganismos e biodiversidade: o novo desafio para a agricultura.** Guaíba, 65-91, 2008.

FISCHBACH, M.A.; WALSH, C.T. **Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanics.** American Chemical Society, Washington, 106: 3468-3496, 2006.

FLEISCHMANN, R.D.; ADAMS, M.D.; WHITE, O.; CLAYTON, R.A.; KIRKNESS, E.F.; KERLAVAGE, A.R.; BULT, C.J.; DOUGHERTY, B.A.; MERRICK, J.M.; MCKENNEY, K.; SUTTON, G.; FITZHUGH, W.; FIELDS, C.; GOCYNE, J.D.; SCOTT, J.; SHIRLEY, R.; LIU, L.; GLODEK, A.; KELLEY, J.M.; WEIDMAN, J. F.; PHILLIPS, C.A.; SPRIGGS, T.; HEDBLUM, E.; COTTON, M.D.; VENTER, J.C. **Whole-genome random sequencing and assembly of haemophilus influenzae Rd.** Science, Washington, 269(52223): 496-512, 1995.

FUNARI, E.; TESTAI, E. **Human health risk assessment related to cyanotoxins exposure.** Critical Reviews in Toxicology, London, 38: 97-125, 2008.

GARCIA-PICHEL, F.; BELNAP, S.N.; SHANZ, F. **Estimates of cyanobacterial biomass and its distribution.** Algological Studies, Stuttgart, 4 (109): 213-227, 2003.

GIGLIO, S.; JIANG, J.; SAINT, C.P.; CANE, D.E.; MONIS, P.T. **Isolation and characterization of the gene associated with geosmin production in cyanobacteria.** Environment, Science and Technology, Washington, 42 (21): 8027-8032, 2008.

GORHAM, P.R.; MCLACHLAN, J.; HAMMER, U.T.; KIM, W.K. **Isolation and culture of toxic strains of *Anabaena flos-aquae* (Lyngb.) de Bréb.** Verhandlungen der Internationalen Vereinigung für Theoretische und Angewandte Limnologie, Berlin, 15: 796-804, 1964.

GRAM, L. **Silent cluster – speak up!** Microbial biotechnology, Malden, 8(1): 13-14, 2014.

GRAMBERGHE, I. VAN.; LELIAERT, F.; MERGEAY, J.; VANORMELINGEN, P.; GUCHT, K.V. der. DEBEER, A.; LACEROT, G.; MEESTER, L.D.; VYVERMAN, W. **Lack of phylogeographic structure in the freshwater cyanobacterium *Microcystis aeruginosa* suggests global dispersal.** Plos One, San Francisco, 6(5): 1-12, 2011.

GRAY, M.W. **Evolution of organellar genome.** Current Opinion in Genetics & Development, Londres, 9: 678–687, 1999.

GRAY, Y.H.M. **It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements.** Trends in genetics, Oxford, 16(10): 461-468, 2000.

GRIESE, M.; LANGE, C.; SOPPA, J. **Ploidy in cyanobacteria.** FEMS Microbiological Letters, Oxford, 323: 124-131, 2011.

HALLEGRAEFF, G.M. **Harmful algal blooms: a global overview.** In: **Manual on Harmful Marine Microalgae.** Hallegraeff, G.M.; Anderson, D.M.; Cembella, UNESCO, Paris, 1-22, 1995.

HARDOIM, C. C. P.; COSTA, R.; ARAÚJO, F. V.; HAJDU, E.; PEIXOTO, R.; LINS, U.; ROSADO, A. S.; VAN ELSAS, J. D. **Diversity of bacteria in the marine sponge *Aplysina fulva* in Brazilian coastal waters.** Applied and Environmental Microbiology, Washington, 75(10): 3331-3343, 2009.

HENDRICKSON, H.; LAWRENCE, J.G. **Selection for chromosome architecture in bacteria.** J. Molecular Biological Evolution, Oxford, 62: 615–629, 2006.

HUMBERT, J; BARBE, V.; LATIFI, A.; GUGGER, M.; CALTEAU, A.; COURSIN, T.; LAJUS, A.; CASTELLI, V.; OZTAS, S.; SAMSON, G., LONGIN, C.; MEDIGUE, C.; MARSAC, N. **A tribute to disorder in the genome of the bloom-forming freshwater cyanobacterium *Microcystis aeruginosa*.** Plos One, San Francisco, 8(8): 1-16, 2013.

ISHIDA, K.; WELKER, M.; CHRISTIANSEN, G.; CODEL-SIX, S.; BOUCHIER, C.; DITTMANN, E.; HERTWECK, C.; MARSAC, N.T. **Plasticity and evolution of aeruginosin biosynthesis in cyanobacteria.** Applied and Environmental Microbiology, Washington, 75(7): 2017-2026, 2009.

JONES, A.C; LIANGCAI, G.; SORRELS, C.M.; SHERMAN, D.H.; GERWICK, W.H. **New tricks from ancient algae: natural products biosynthesis in marine cyanobacteria.** Current Opinion In Chemical Biology, Londres, 13: 216-223, 2009.

KAEBERNICK, M.; NEILAN, B.A. **Ecological and molecular investigations of cyanotoxin production.** FEMS Microbiology Ecology, Oxford, 35: 1-9, 2001.

KANEKO, T.; SATO, S.; KOTANI, H.; TANAKA, A.; ASAMIZU, E.; NAKAMURA, Y.; MIYAJIMA, N.; HIROSAWA, M.; SUGIURA, M.; SASAMOTO, S.; KIMURA, T.; HOSOUCHI, T.; MATSUNO, A.; MURAKI, A.; NAKAZAKI, N.; NARUO, K.; OKUMURA, S.; SHIMPO, S.; TAKEUCHI, C.; WADA, T.; WATANABE, A.; YAMADA, M.; YASUDA, M.; TABATA, S. **Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions.** DNA Research, Oxford, 3(3): 109–136, 1996.

KAPLAN, A.; HAREL, M.; KAPLAN-LEVY, R. N.; HADAS, O.; SUKENIK, A.; DITTMAN, E. **The languages spoken in the water body (or the biological role of cyanobacterial toxins).** Frontiers in Microbiology, Lausanne, 3 (138), 2012. doi: 10.3389/fmicb.2012.0013

KEHR, J-C.; PICCHI, D.G.; DITTMANN, E. **Natural product biosyntheses in cyanobacteria: A treasure trove of unique enzymes.** Beilstein Journal of Organic Chemistry, Beilstein, 7: 1622-1635, 2011.

KOONIN, E.V. **Evolution of genome architecture.** The International Journal of Biochemistry & Cell Biology, Londres, 41: 298–306, 2009.

KOONIN, E.V. **Are there laws for genome evolution?** Plos Computational Biology, San Francisco, 7(8): 1-6, 2011. e1002173

KUNIN, V.; OUZOUNIS, C.A. **The balance of driving forces during genome evolution in prokaryotes.** Genome Research, Cold Spring Harbor, 13:1589–1594, 2003.

KRAUZE, D.J.; DIDELOT, X.; CADILLO-QUIROZ, H.; WHITAKER, R.J. **Recombination shapes genome architecture in an organism from the Archeal domain.** *Genome Biology and Evolution*, Oxford, 6(1): 170-178, 2013.

KREITLOW, S.; MUNDT, S.; LINDEQUIST, U. **Cyanobacteria – a potential source of new biologically active substances.** *Journal of Biotechnology*, Londres, 70: 61-63, 1999.

LEIKOSKI, N.; LIU, L.; JOKELA, J.; WAHLSTEN, M.; GUGGER, M.; CALTEAU, A.; PERMI, P.; KERFELD, C.A.; SIVONEN, K.; FEWER, D.P. **Genome mining expands the chemical diversity of the cyanobactin family to include highly modified linear peptides.** *Chemistry & Biology*, Londres, 20: 1033-1043, 2013.

LIN, S.; HAAS, S.; ZEMOJTEL, T.; XIAO, P.; VINGRON, M.; LI, E. **Genome-wide comparison of cyanobacterial transposable elements, potential genetic diversity indicators.** *Gene*, Londres, 473(2): 139-149, 2011.

MARAHIEL, M.A.; STACHELHAUS, T.; MOOTZ, H.D. **Modular peptide synthetases in nonribosomal peptide synthesis.** *Chemical Reviews*, Washington, 97(7): 2651-2673, 1997.

MEDEMA, M.; BLIN, K.; CIMERMANCIC, P.; JAGER, V. de; ZAKRZEWSKI, P.; FISCHBACH, M.A.; WEBER, T.; TAKANO, E.; BREITLING, R. **antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequence.** *Nucleic Acid Research*, Oxford, 39: 1-8, 2011.

MICALLEFF, M.L.; D'AGOSTINO, P.; AL-SINAWIB, B.; NEILAN, B.A.; MOFFITT, M.C. **Exploring cyanobacterial genomes for natural product biosynthesis pathways.** *Marine Genomics*, Londres, 2014. doi: 10.1016/j.margen.2014.11.009

NAMIKOSHI, M.; RINEHART, K.L. **Bioactive compounds produced by cyanobacteria.** *Journal of Industrial Microbiology & Biotechnology*, Londres, 17: 373-384, 1996.

NELSON D.L.; COX, M.M. **Principles of biochemistry**. W. H. Freeman and Company, Nova Iorque, 5^a ed. 1302 pp, 2008.

NUNNERY, J.K.; MEVERS, E.; GERWICK, W.H. **Biologically active secondary metabolites from marine cyanobacteria**. Current Opinion in Biotechnology, Londres, 21:787–793, 2010.

PAERL, H.W.; OTTEN, T.G. **Harmful cyanobacterial blooms: causes, consequences, and controls**. Microbial Ecology, Londres, 65(4): 995-1010, 2013.

PEARSON, L.; MIHALI, T.; MOFFITT, M.; KELLMANN, R.; NEILAN, B. **On the chemistry, toxicology and genetics of the cyanobacterial toxins, microcystin, nodularin, saxitoxin and cylindrospermopsin**. Marine Drugs, 8: 1650-1680, 2010.

OCHMAN, H.; DAVALOS, L. **The Nature and Dynamics of Bacterial Genomes**. Science, Washington, 311: 1730-1733, 2006.

ROUHIAINEN, L.; VAKKILAINEN, T.; SIEMER, B.L.; BUIKEMA, W.; HASSELKORN, R.; SIVONEN, K. **Genes coding for hepatotoxic heptapeptides (Microcystins) in the cyanobacterium *Anabaena* strain 90**. Applied and Environmental Microbiology, Washington, 70: 686-692, 2004.

ROUNGE, T.B.; ROHRLACK, T.; NEDERBRAGT, A.J.; KRISTENSEN, T.; JAKOBSEN, K.S. **A genome wide analysis of nonribosomal peptide synthetase gene clusters and their peptides in a *Planktothrix rubescens* strain**. BMC Genomics, Londres, 10: 396, 2009.

SABART, M.; POBEL, D.; LATOUR, D.; ROBIN, J.; SALENÇON, M.J.; HUMBERT. **Spatiotemporal changes in the genetic diversity in French bloom-forming populations of the toxic cyanobacterium, *Microcystis aeruginosa***. Environmental Microbiology Reports, Hoboken, 1(4): 263-272, 2011.

SANT'ANNA, C.L.; CARVALHO, L.R.; FIORE, M.F.; SILVA-STENICO, M.E.; LORENZI, A.S.; RIOS, F.R.; KONNO, K.; GARCIA, C.; LAGOS, N. **Highly toxic *Microcystis aeruginosa* strain, isolated from São Paulo-Brazil, produce hepatotoxins and paralytic shellfish poison neurotoxin.** Neurotoxicity Research, Washington, 19: 389-402, 2011.

SANTOS, K. R. S.; SANT'ANNA, C. L. **Cianobactérias de diferentes tipos de lagoas (“salina”, “salitrada” e “baía”) representativas do Pantanal da Nhecolândia, MS, Brasil.** Revista Brasileira de Botânica, São Paulo, 33(1):. 61-83, 2010

SASSO, S.; POHNERT, G.; LOHR, M.; MITTAG, M.; HERTWECK, C. **Microalgae in the postgenomic era: a blooming reservoir for new natural products.** FEMS Microbiology Reviews, Amsterdam, 36(4): 761-785, 2011.

SCHOPF, J.W. **The fossil record: tracing the roots of the cyanobacterial lineage.** In: Whitton, B.A.; Potts, M. (Ed.). **The ecology of cyanobacteria.** Dordrecht: Kluwer Academic, Dordrecht, 13-35, 2000.

SCHWARTZER, D.; FINKING, R.; MARAHIEL, M. **Nonribosomal peptides: from genes to products.** Natural Products Reports, Cambridge, 20: 275–287, 2002.

SHEN, B. **Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms.** Current Opinion in Chemical Biology, Londres, 7: 285-295, 2003.

SKULBERG, O.M. **Biophotolysis, hydrogen production and algal culture technology.** In: Yürüm, Y. **Hydrogen energy system. Production and utilization of hydrogen and future aspects.** Dordrecht: Kluwer Academic, Dordrecht, 95-110, 1995.

STEWART, W.D.P. **Nitrogen fixation.** Botanical Monographs, Laramie, 9: 260-278, 1973.

SUKENIK, A.; KAPLAN-LEVY, R.; WELCH, J.M.; POST, A.F. **Massive multiplication of genome and ribosomes in dormant cells (akinetes) of *Aphanizomenon ovalisporum* (Cyanobacteria).** Internacional Society for Microbial Ecology, Wageningen, 6: 690-679, 2012.

SUMMONS, R.E.; JAHNKE, L.L.; HOPE, J.M.; LOGAN, G.A. **2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis.** Nature, Londres, 400: 554-557, 1999.

SUURNÄKKI, S.; GOMEZ-SAEZ, G.V.; RANTALA-YLINEN, A.; JOKELA, J.; FEWER, D.P.; SIVONEN, K. **Identification of geosmin and 2-methylisoborneol in cyanobacteria and molecular detection methods for the producers of these compounds.** Water Research, Londres, 68: 56-66, 2014.

TILLET, D.; DITTMANN, E.; ERHARD, M.; VON DÖHREN, H.; BÖRNER, T.; NEILAN, B.A. **Structural organization of microcystin biosynthesis in *Microcystis aeruginosa* PCC7806: an integrated peptide-polyketide synthetase system.** Chemical Biology, Washington, 7(10): 753-764, 2000.

TOMITANI, A.; KNOLL, A. H.; CAVANAUGH, C. M.; OHNO, T. **The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives.** Proceedings of the National Academy of Sciences of the USA, Washington, 103(14): 5442-5447, 2006.

VARGAS, M.; KASHIFI, K.; BLUNT-HARRIS, E.L.; LOVLEY, D.R. **Microbial evidence for Fe(III) reduction on early Earth.** Nature, Londres, 395(6697): 65-67, 1998.

VON DÖHREN, H.; KELLER, U.; VATER, J.; ZOCHER, R. **Multifunctional peptide synthetases.** Chemical Reviews, Washington, 29(7): 2675-2706, 1997.

WATSON, J.D.; BAKER, T.A.; BELL, A.P.; GANN, A.; LEVINE, M.; LOSICK, R. **Molecular biology of the gene.** Benjamin Cummings, Cold Spring Harbor, 5^a ed. , 755 pp, 2004.

WELKER, M.; VON DÖHREN, H. **Cyanobacterial peptides – nature’s own combinatorial biosynthesis**. FEMS Microbiology Reviews, Amsterdam, 30(4): 530-563, 2006.

WHITTON, B.A.; POTTS, M. Introduction to cyanobacteria. *In*: Whitton, B.A.; Potts, M. (Ed.). **The ecology of cyanobacteria**. Dordrecht: Kluwer Academic, Dordrecht, 2000.

WILSON, M.W.; MORI, T.; RUCKERT, C.; URIA, A.; HELF, M.J.; TAKADA, K.; GERNERT, C.; STEFFENS, U.A.E.; HEYCKE, N.; SCHMITT, S.; RINKE, C.; HELFRICH, E.J.N.; BRACHMANN, A.O.; GURGUI, C.; WAKIMOTO, T.; KRACHT, M.; CRUSEMANN, M.; HETSCHER, U.; ABE, I.; MATSUNAGA, S.; KALINOWSKY, J.; TAKEYAMA, H.; PIEL, J. **An environmental bacterial taxons with a large and distinct metabolic repertoire**. Nature, Londres, 1-19, 2014.

WEBER, T.; BLIN, K.; DUDELA, S.; KRUG, D.; KIM, H.U.; BRUCCOLERI, R.; LEE, S.Y.; FISCHBACH, M.A.; MÜLLER, R.; WOHLLEBEN, W.; BREITLING, R.; TAKANO E.; MEDEMA, M.H. **antiSMASH 3.0 — a comprehensive resource for the genome mining of biosynthetic gene clusters**. Nucleic Acids Research, Oxford, 43: 237-243, 2015.

WINTER, JM; BEHNKEN, S.; HERTWECK, C. **Genomics-inspired discovery of natural products**. Current Opinion in Chemical Biology, Londres, 15(1): 22-31, 2011.

XU, H.; PAERL, H.W.; QUIN, B.; ZHU, G.; GAO, G. **Nitrogen and phosphorous inputs control phytoplankton growth in eutrophic Lake Taihu, China**. Limnology and oceanography, Weinheim, 55(1): 420-432, 2010.

ZIEMERT, N.; ISHIDA, K.; PHILIPPE, Q.; BOUCHIER, C.; HERTWECK, C.; MARSAC, N.T.; DITTMANN, E. **Microcyclamide Biosynthesis in Two Strains of *Microcystis aeruginosa*: from Structure to Genes and Vice Versa**. Applied and Environmental Microbiology, Londres, 74(6): 1791-1897, 2008.