TITLE: ANNOTATION AND GENOME MINING OF *GEMINOCYSTIS* GBBB08 STRAIN FROM CERRADO BIOME

AUTHORS: BARROS, J.I.P.[1]; COLINS, M.S.[1]; COSTA, A.L.S.[1]; RIBEIRO, I.S.[1]; BUTARELLI, A.C.A.[2]; FERREIRA, L.S.S.[2]; DALL'AGNOL, H.[1]; DA SILVA, A.M.[2]; SETUBAL, J.C.[2]; DALL'AGNOL, L.T.[1]

INSTITUTIONS: [1]UNIVERSIDADE FEDERAL DO MARANHÃO, SÃO LUÍS, MA (AVENIDA DOS PORTUGUESES, 1966, CEP 65080-805, SÃO LUÍS – MA, BRAZIL). [2]UNIVERSIDADE DE SÃO PAULO, SÃO PAULO, SP (Av. Prof. Lineu Prestes, 748 - Butantã, São Paulo - SP, 05508-900, SÃO PAULO – SP, BRAZIL).

ABSTRACT:

Geminocystis is a genus of cyanobacteria described in 2009 based on studies of the genus *Synechocystis*, which still has little data available about its genomic and metabolic diversity, especially regarding its hypothetical proteins. Manual annotation and curation make it possible to identify and detail with high reliability the hypothetical proteins, distinctly classified by their functions and structures. Therefore, the objective of this work was to carry out the annotation and in silico analysis of the genome of the non-axenic *Geminocystis* strain GBBB08 originally isolated from the Cerrado biome and deposited at the GBBB collection at UFMA and sequenced at the Center for Advanced Technologies in Genomics at Chemistry Institute from USP. After metagenomic sequencing, de novo assembly was performed with the Metawrap Assembly module using metaSPAdes v. 3.13. Assembled contigs were subjected to three different binning rounds using CONCOCT, MaxBin2, and MetaBAT2. The annotation of the isolated genome of the GBBB01 strain was performed using the Prokaryotic Genome Annotation Pipeline (PGAP) and RAST. In the structural and comparative characterization of the genome, antiSMASH 6.0, NaPDoS, PHASTER, CRISPRCasFinder, TYGS, GGDC and MEGA X tools were used. Manual curation of the annotation of gene clusters was performed using the computational tool BLAST v2.6.0+ (NCBI), PFAM, UniProtKB, HAMAP and its visualization by Artemis. The analysis of hypothetical proteins targeted their physicochemical properties and sub-cellular localization, using sequence comparison by BLAST in ProtParam, PSORTb, GRAVY and STRING databases. We identified 13 biosynthetic gene clusters related to the production of anabaenopeptin/nostamide, terpenes, betalactones, heptadecene, RiPP recognition element and other types of NRPS/PKS pathways. Within the biosynthetic pathways identified by antiSMASH it was possible to select 32 hypothetical proteins of which 12 HC-HPs were classified with high confidence and had their annotation optimized. The present work enabled a better understanding of the metabolic and ecological potential of this strain through a systematic and integrated approach, and consequently allowing a better evaluation of the biodiversity of this region and its value.

Keywords: Cyanobacteria, Hypothetical proteins, in silico.

Development agency: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).