Feature-based molecular networking and *in silico* structure annotation/classification of LC-HRMS data unfold the chemodiversity of cyanobacteria in environmental samples.

ILIAKOPOULOU S.[1,2], ZERVOU S.K.[1], TRIANTIS T.M.[1], HISKIA A.[1], KALOUDIS T.[1,3*]

[1]PHOTOENV Lab, Institute of Nanoscience and Nanotechnology (INN), NCSR Demokritos, Athens, Greece
[2]Department of Environmental Engineering, University of Patras, Greece
[3]Laboratory of Organic Micropollutants, Water Quality Control Department, Athens Water Supply and Sewerage Company (EYDAP SA), Greece

*corresponding author:
t.kaloudis@inn.demokritos.gr , kaloudis@eydap.gr

Abstract Freshwater cyanobacteria are prominent sources of structurally diverse natural compounds. Bioactive cyanometabolites are particularly relevant to water quality and public health protection. Non-targeted analysis (NTA) by liquid chromatography - high resolution mass spectrometry (LC-HRMS) is applied to expand the range of detected and identified metabolites. However, data analysis is challenging and subjected to limitations arising from the availability of experimental or library-based mass spectra. We present an HRMS data analysis workflow using state-of-the-art computational tools that we have applied to analyze samples from cyanobacteria blooms in Greek lakes. Pre-processing of data was carried out in MZmine3 (feature detection, deconvolution, alignment, deisotoping, gap filling). Processed data were exported in GNPS for feature-based molecular networking - FBMN and annotations based on public GNPS libraries. In parallel, feature lists were processed in SIRIUS and its associated tools, for de novo molecular formula annotation, database search, prediction of compound classes using molecular fingerprints, and ranking of candidates using fragmentation trees. Results were visualized and further explored in Cytoscape, to enable annotation propagation. Such workflows substantially expand the chemical space of annotated cyanometabolites at structural and compound-class levels, and the discovery of new compounds which are not included in libraries.

Keywords: Cyanobacteria metabolites, MZmine3, GNPS, SIRIUS, LC-HRMS

Introduction

Comprehensive studies of the metabolism of free-living organisms obtained from the natural environment and their responses to natural and anthropogenic stressors, in the context of environmental metabolomics, contribute to risk assessment for the protection of human health and the environment [1]. Cyanobacteria are a source of a plethora of secondary metabolites in freshwater bodies, with more than 2000 compounds characterized so far [2]. Exploration of the metabolic potential of cyanobacteria is promoted by advances in High Resolution Mass Spectrometry (HRMS), modern computational tools, spectral libraries and databases [3]. Especially, untargeted analysis using LC-HRMS in Data Dependent Acquisition mode (DDA) widely extends the numbers of detected and annotated metabolites [4]. Modern computational platforms and tools aim to increase the number of detected and annotated cyanobacteria metabolite annotations at structural- and compound class- levels [3]. Here, we present a LC-DDA-HRMS computational workflow and its potential in substantially expanding annotations of cyanobacteria metabolites in environmental samples, as shown for cyanobacteria bloom samples of lake Marathonas, Greece.

Methods

Samples from cyanobacteria blooms in lake Marathonas were collected on 26 October 2010 and 28 September 2020. Samples were filtered (45mm glass fiber filters), lyophilized (Alpha 3-4, LSCbasic, Martin Christ) and stored at -20 ºC. A standard solution of 14 cyanopeptides and blank samples were analyzed in parallel for quality control. Extraction (75% methanol) and centrifugation prior to analysis was carried out as described before [5].
Samples were analyzed by HPLC (Ultimate 3000 RSLC, ThermoFisher Scientific) coupled to HRMS (Orbitrap Fusion Lumos Tribrid, ThermoFisher Scientific). An Atlantis T3 column was used (2.1 mm x 100 mm, 3 μm, Waters) with gradient elution (acetonitrile/water with 0.5% formic acid) and a flow of 0.2 ml/min. A heated electrospray source (HESI) was used in positive polarity. Acquisition of MS data (DDA) was carried out by MS1 scans (m/z 160 – 2000), and subsequent fragmentation by CID και HCD using dynamic exclusion.

Results

A workflow for MS data processing and analysis was developed (Figure 1). Raw MS data (.raw) were converted to mzML using MSconvert [6] and then pre-processed in MZmine3 [7] for mass detection (MS1 and MS2), peak picking, deconvolution, feature alignment, deisotoping and gap-filling. Exports from MZmine3 (feature table, MGF spectra file) and a metadata file, were used for feature-based molecular networking (FBMN) in GNPS [8]. Mass spectra (mgf format) were imported in SIRIUS [9] using CSI:FingerID [10] with COSMIC [11], ZODIAC [12] and CANOPUS [13], for de novo molecular formula annotation, database search, prediction of compound classes using molecular fingerprints, and ranking of candidates using fragmentation trees. The molecular network and combined annotations from GNPS and SIRIUS were visualized in Cytoscape [14].

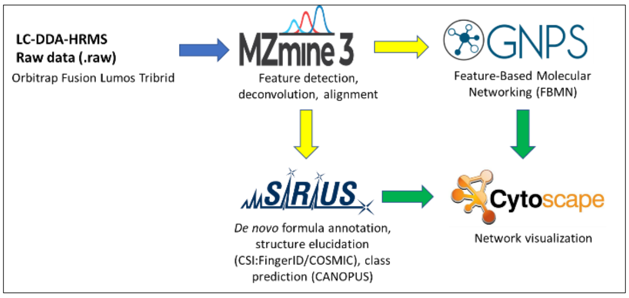
LC-DDA-HRMS
Raw data (.raw)
Orbitrap Fusion Lumos Tribrid
MZmine 3
Feature detection, deconvolution, alignment
GNPS
Feature-Based Molecular Networking (FBMN)
SIRIUS
De novo formula annotation, structure elucidation (CSI:FingerID/COSMIC), class prediction (CANOPUS)
Cytoscape
Network visualization

Figure 1. MS data analysis workflow

As an example, a total of 2189 MS features were extracted from a dataset of the two bloom samples, a blank and a reference standard. The molecular network included 35 subnetworks and 218 singleton nodes. Twenty-four nodes were annotated by GNPS libraries (mass error <0.02 Da) and additional nodes were annotated by Delta m/z and examination of mass spectra within subnetworks. An example of a subnetwork of microcystins is shown in Figure 2.

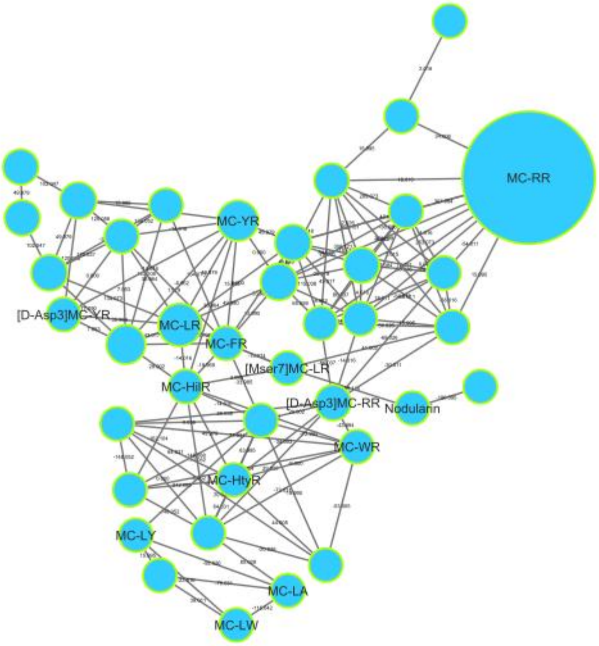

Figure 2: A subnetwork of microsystins.

In addition, SIRIUS annotated more than 50 nodes in superclasses such as oligopeptides, glycerolipids, glycerophospholipids, nucleosides.

The combination of LC-HRMS with feature-based molecular networking and *in silico* machine-learning tools enables the propagation of annotations within the networks, which considerably widens the scope of studies of cyanobacteria metabolites. We are currently using this approach for comprehensive studies of metabolites, natural and transformation products in cyanobacteria blooms in lakes of Greece.

Acknowledgements

We warmly thank the developers of GNPS, MZmine3, SIRIUS, Cytoscape (refs. 6-14) for maintaining these open platforms and tools for the benefit of the community.

We are grateful to Prof. Ryszard Lobinski, Prof. Joanna Szpunar and Dr. Simon Godin of IPREM-CNRS, Univ. of Pau, for providing us access to the LC-HRMS system used in this study.

References

1. Marion G. Miller (2007). Environmental Metabolomics: A SWOT Analysis (Strengths, Weaknesses, Opportunities, and Threats). Journal of Proteome Research 2007 6 (2), 540-545.

2. Martin R. Jones, Ernani Pinto, Mariana A. Torres et al., CyanoMetDB, a comprehensive public database of secondary metabolites from cyanobacteria, Water Research, Volume 196, 2021, 117017.

3. De Jonge NF, Mildau K, Meijer D, Louwen JJR, Bueschl C, Huber F, van der Hooft JJJ. Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. Metabolomics. 2022 Dec 5;18(12):103.

4. Fabio Varriale, Luciana Tartaglione, Sevasti-Kiriaki Zervou, et al. (2022). Untargeted and targeted LC-MS and data processing workflow for the comprehensive analysis of oligopeptides from cyanobacteria. Chemosphere, 137012.

5. Zervou, S.-K., Gkelis, S., Kaloudis, T., Hiskia, A., Mazur-Marzec, H. (2020). New microginins from cyanobacteria of Greek freshwaters. *Chemosphere* 248, 125961.

6. Chambers, M., Maclean, B., Burke, R. et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30, 918–920 (2012).

7. Schmid, R., Heuckeroth, S., Korf, A. et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. Nature Biotechnology (2023).

8. Nothias, L.-F., Petras, D., Schmid, R. et al. Feature-based molecular networking in the GNPS analysis environment. Nat. Methods 17, 905–908 (2020).

9. Kai Dührkop, Markus Fleischauer, Marcus Ludwig, Alexander A. Aksenov, Alexey V. Melnik, Marvin Meusel, Pieter C. Dorrestein, Juho Rousu, and Sebastian Böcker, SIRIUS 4: Turning tandem mass

spectra into metabolite structure information. Nature Methods 16, 299–302, 2019.

10. Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu, and Sebastian Böcker. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proceedings of the National Academy of Sciences U S A 112(41), 12580-12585, 2015.
11. Hoffmann, M.A., Nothias, LF., Ludwig, M. et al. High-confidence structural annotation of metabolites absent from spectral libraries. Nat Biotechnol 40, 411–421 (2022).
12. Ludwig, M., Nothias, LF., Dührkop, K. et al. Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. Nat Mach Intell 2, 629–641 (2020).
13. Kai Dührkop, Louis-Félix Nothias, Markus Fleischauer, et al.. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. Nature Biotechnology, 39, 462–471 (2021).
14. Shannon P, Markiel A, Ozier O, et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks". Genome Res. 13 (11): 2498–504.