ARPHA Conference Abstracts 4: e65047
doi: 10.3897/aca.4.e65047

Conference Abstract

Strategies for the bioinformatic treatment of high-throughput sequencing data for diatom studies and bioassessment

Kálmán Tapolczai[‡,§], François Keck[|], Valentin Vasselon[¶], Géza B. Selmeczy[#], Maria Kahlert[¤], Agnès Bouchez[«], Frédéric Rimet[»], Judit Padisák[#]

‡ University of Pannonia, Veszprém, Hungary
§ Hungarian Academy of Sciences (MTA), Budapest, Hungary
| Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
¶ SCIMABIO-Interface, Thonon-les-Bains, France
# University of Pannonia, Center for Natural Science, Research Group of Limnology, Veszprém, Hungary
¤ Swedish University of Agricultural Sciences, 750 07 Uppsala, Sweden
« INRAE, USMB, UMR CARRTEL, Thonon-les-bains, France
» INRAE, USMB, UMR CARRTEL, Thonon-les-Bains, France

Corresponding author: Kálmán Tapolczai (tapolczai.kalman@gmail.com)

Received: 25 Feb 2021 | Published: 04 Mar 2021

Citation: Tapolczai K, Keck F, Vasselon V, Selmeczy GB, Kahlert M, Bouchez A, Rimet F, Padisák J (2021) Strategies for the bioinformatic treatment of high-throughput sequencing data for diatom studies and bioassessment. ARPHA Conference Abstracts 4: e65047. https://doi.org/10.3897/aca.4.e65047

Abstract

Diatom biomonitoring and ecological studies can greatly benefit from DNA metabarcoding compared to conventional microscopical analysis by potentially providing more reliable and accurate data in a cost- and time-efficient way. A conventional strategy for the bioinformatic treatment of sequencing data involves the clustering of quality filtered sequences into Operational Taxonomic Units (OTUs) based on a global sequence similarity, and their assignment to taxonomy using a reference library. Then, the obtained species lists of the successfully assigned taxa are used for subsequent analyses or quality index calculation.

However, the high diversity of bioinformatic methods and parameters make inter-studies comparison difficult, especially because OTUs are specific to a given study. Clustering sequences into OTUs aims to reduce the biasing effect of sequencing artefacts and to reach an approximate species level delimitation at the price of potentially grouping together sequences with different ecology. A similar bias occurs when sequences that differ from each other by their ecological preference are assigned to the same taxa. The incompleteness of reference libraries can further introduce a bias by not taking into account unassigned sequences, thus losing the ecological information they possess.

In order to overcome these biases, our studies tested new approaches on *de novo* developed diatom indices based on periphytic samples collected from streams in France and Hungary. Index development was performed with the *leave-one-out cross validation* (LOOCV) technique by building a model on a training dataset containing *n-1* samples and testing it on the remaining test sample. Test values were correlated with a reference environmental gradient. The model was based on the calculation of optimum and tolerance of taxonomic units along the reference gradient and a modified Zelinka-Marvan diatom index equation. Taxonomic units tested in the studies were morphospecies, OTUs (95% similarity threshold), Individual Sequence Units (ISUs, via minimal bioinformatic quality filtering) and Exact Sequence Variants (ESVs, via DADA2 denoising algorithm).

The “clustering-free” approach (ISU- and ESV-based indices) performed better than the OTU-based one, providing a fine taxonomic resolution where the ecological difference on genetically close sequence variants could be detected. Thus, these indices are more adapted to a standardized and comparable routine bioassessment. The “taxonomy-free” approach revealed the ecological preferences for those molecular taxonomic units (ISUs/ESVs) that otherwise either (i) would have been assigned to the same taxa due to genetic similarity, or (ii) would not have been recognized because of their absence from the reference libraries. However, we also found that taxonomic information cannot be neglected in ecological studies when the presence of organisms under particular environmental conditions is to be explained or interpreted e.g. via the traits they possess. New types of clustering methods are welcome in the future of biomonitoring where the delimitation of taxonomic units should be refined based on a higher emphasis on their ecology rather than on morphological or genetical criteria.

Keywords

bioassessment, biomonitoring, diatoms, DNA metabarcoding, ESV, ISU, OTU, taxonomy-free

Presenting author

Kálmán Tapolczai

Presented at

1st DNAQUA International Conference (March 9-11, 2021)