Paper submitted to GigaByte and dataset to GigaDB

Title: A microbe taxa list dataset and template to assess the ecological status of marine sediments and waters

Author: Angel Borja (Orcid: 0000-0003-1601-2025)

Affiliation: AZTI, Marine Research, Basque Research and Technology Alliance (BRTA), Pasaia, Spain

Abstract

Microbes have usually been neglected as indicators to assess the ecological status, under multiple human pressures. Some years ago, a biotic index (microgAMBI) was proposed to assess the status of marine sediments and waters, and it has been tested under different pressures and biogeographical areas. The index is based on the assignation of each taxon to an ecological group (sensitive or not to disturbance), and this list has grown since the first publication. Because the increasing use of the index, it could be wise to make available the calculation template and the updated taxa list (1,969 taxa currently), meeting the FAIR (Findable, Accessible, Interoperable and Reusable) principles. This manuscript describes the dataset and template.

1.- Data description

1.1. Context

Assessing the ecological status of different biological elements is necessary to take informed management decisions, to reduce the effects of multiple pressures at sea (Birk et al., 2020; Korpinen et al., 2021). Although many biotic indices have been proposed to assess the ecological status of phytoplankton, macroalgae, seagrasses, macroinvertebrates or fish (Birk et al., 2012), microbes have been usually neglected as elements to assess the marine ecological status, and only recently they have been investigated in response to anthropogenic disturbances (Ager et al., 2010). However, some years ago, Aylagas et al. (2017) developed a taxonomy-based biotic index (microgAMBI), using bacterial community composition data, for marine sediment assessment. This index is based on the principles of AZTI's Marine Biotic Index (AMBI), developed by Borja et al. (2000), to assess the status of benthic macroinvertebrate communities. AMBI is based on the assignation of each species to an ecological group (EG) (i.e. EG I, sensitive species; EG II, indifferent; EG III, tolerant; EG IV, second-order opportunistic species; and EG V, first order opportunistic species). This was further developed as gAMBI (genomic AMBI), in which species are identified using metabarcoding (Aylagas et al., 2014).

MicrogAMBI is calculated from 16S rRNA metabarcoding data, in both coastal and estuarine locations. Originally developed in the north of Spain, it was validated against a pressure index measuring the anthropogenic disturbance (Aylagas et al., 2017). It has been further applied in multiple biogeographical areas across the ocean (polar, tropical, temperate), including water column, sediment and corals, as well as different pollution sources, which include wastewater discharge, eutrophication, hydrocarbon concentration, aquaculture (Borja, 2018; Pearman et al., 2018; Clark et al., 2021; Aylagas et al., 2021; Sun et al., 2021; Lanzén et al., 2021; Bourhane et al., 2022; etc.).

The initial paper included a list of around 800 taxa, which have been increased with posterior publications, in which authors provided the information on the EGs used. However, as the list has grown until near 2,000 taxa, it is necessary to make it available, meeting the FAIR (Findable, Accessible, Interoperable and Reusable) principles. This is why this dataset is made available through *GigaByte* and the *GigaDB* repository.

1.2. Methods

Microbial taxa are identified via the taxonomic classification of sequences and then assigned to ecological groups based on surveyed literature, as done in the initial paper (Aylagas et al., 2017). Following the principle of AMBI, but adapted to two single responses, those

microbes that are not associated with pollution inputs, are included as sensitive and indifferent taxa (EGI) and those associated with pollution inputs are included as tolerant and opportunistic taxa (EGIII) (Figure 1). The ratio of the relative abundance of EGI and EGIII in a sample is used to calculate the index which in turn provides an ecological classification of, high, good, moderate, poor, or bad status (Aylagas et al., 2017) (Figure 1).

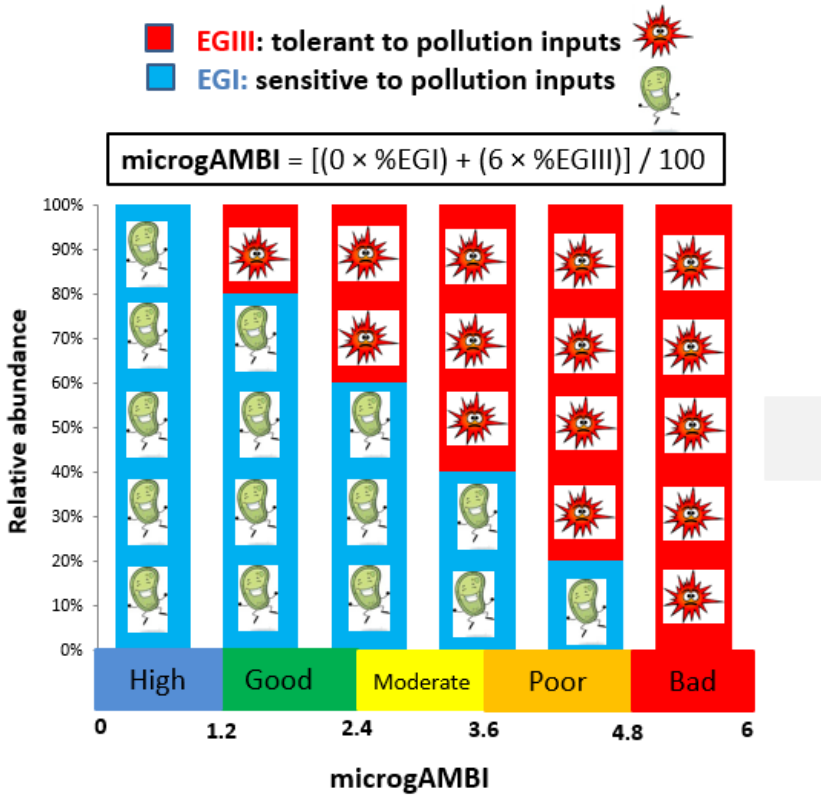

Figure 1: MicrogAMBI calculation, showing the relative abundance of the taxa assigned to each of the two ecological groups (EGI, sensitive to pollution; and EGIII tolerant), the equation to calculate the index, and the boundary values determining the five quality classes (modified and updated from Aylagas (2017), based on the information from Aylagas et al. (2017) and Borja (2018).

To expand the initial taxa list, a review of the literature was done to select common microbial taxa appearing in environmental studies. The assignation criteria to the above two EGs, are the same as in Aylagas et al. (2017). A taxon is assigned to EGIII when (i) dominates in organic matter-enriched sediments; (ii) presents organic pollution response; (iii) dominant presence in anoxic methane-rich sediments; (iv) identified as nitrite oxidizer and related to nitrogen inputs; (v) present in sulfide-rich wastewaters; (vi) present in wastewater treatment plants; (vii) has a role in methanogenic degradation of alkanes; (viii) has a role in aromatic compounds biodegradation, including petroleum products pollution, as complex PAHs; and (ix) is a potential pathogen. The remainder taxa are assigned to EGI (e.g. aerobic taxa, taxa described

as living in pristine systems, etc.). Taxa of unknown ecological function are included as 'not assigned'.

1.3. Data validation and quality control

The taxa list includes, for each taxon, the reference or web page link in which the assignation is based. This information is available in the database in *GigaDB*, as a template allowing the calculation of microgAMBI (microgAMBI-template-version-2023-04-22.xls).

This template has three sheets. The first one ('readme'), includes a guide with seven steps on how the user should prepare the data to calculate the index. The second one ('template'), includes the empty template to be filled in with the data and the necessary equations to obtain the ecological status by station, as described initially in Aylagas et al. (2017). The third one ('taxa list'), includes: (i) in column A, the correlative number of each taxon; (ii) in column B, the name of each taxon (in the case of genus, without 'sp.', hence, it is better to remove this in the datasets to be tested); (iii) in column C, the assignation to the corresponding EG; and (iv) in column D, the literature in which the decision to assign a taxon to an EG is based on.

The current list includes 1,969 taxa (820 in EGI; 1,124 in EG III; and 25 as not assigned), for each of them, one or several papers to support the assignation decision are included in the list; in other cases, it could be a web page in which the decision is based; in some cases, it could be based by analogies with other species within the same genus.

To ensure the quality control, the database is public, and any researcher can suggest adding new taxa or change an assignation, based on new evidence to support such change, by contacting the author of this paper (aborja@azti.es) or a member of his team, as included in the 'readme' sheet of the dataset in *GigaDB*. Meeting the FAIR principles, allow future updating of the taxa list and the template, as well as the reuse of this dataset.

2.- Re-use potential

Any author having a set of microbial data, based on metabarcoding, with the number of reads by station sampled, can easily use this template to calculate the ecological status in the study area. Having gradients of impact (e.g. a wastewater discharge, and aquaculture farm, an industrial activity with contaminants, etc.) could be useful to better interpret the results. For doing that, simply, follow the instructions in the readme sheet, copy and paste your data and look at the microgAMBI results. Examples of such applications can be seen in Borja (2018),

Pearman et al. (2018), Clark et al. (2021), Aylagas et al. (2021), Sun et al. (2021), Lanzén et al. (2021), and Bourhane et al. (2022), among others. If a high percentage of taxa is not assigned, the authors using the template can contact the author of the database and try to find together evidence on the ecological group to assign the taxa, updating the database with new taxa, which can be useful for the users community.

3.- Data Availability

Data is available in the GigaScience GigaDB repository (DOI: xxxxx).

4.- Funding

This manuscript is a result of GES4SEAS (Achieving Good Environmental Status for maintaining ecosystem services, by assessing integrated impacts of cumulative pressures) project, funded by the European Union under the Horizon Europe program (grant agreement no. 101059877), www.ges4seas.eu.

5.- Author contribution

The author has completed the dataset, maintained and curated it, as well as written the supporting paper.

6.- Acknowledgements

Eva Aylagas (King Abdullah University for Science and Technology) and Andres Lanzén (AZTI) have provided ideas and inputs through the collation process.

7.- References

Ager, D., S. Evans, H. Li, A. K. Lilley, C. J. Van Der Gast, 2010. Anthropogenic disturbance affects the structure of bacterial communities. Environmental Microbiology, 12: 670-678.

Aylagas, E., 2017. DNA metabarcoding derived biotic indices for marine monitoring and assessment. PhD Thesis Dissertation, University of the Basque Country, 248 pp.

Aylagas, E., Á. Borja, N. Rodríguez-Ezpeleta, 2014. Environmental Status Assessment Using DNA Metabarcoding: Towards a Genetics Based Marine Biotic Index (gAMBI). Plos ONE, 9: e90529.

Aylagas, E., Á. Borja, M. Tangherlini, A. Dell'Anno, C. Corinaldesi, C. T. Michell, X. Irigoien, R. Danovaro, N. Rodríguez-Ezpeleta, 2017. A bacterial community-based index to assess the ecological status of estuarine and coastal environments. Marine Pollution Bulletin, 114: 679-688.

Aylagas, E., J. Atalah, P. Sánchez-Jerez, J. K. Pearman, N. Casado, J. Asensi, K. Toledo-Guedes, S. Carvalho, 2021. A step towards the validation of bacteria biotic indices using DNA metabarcoding for benthic monitoring. Molecular Ecology Resources, 21: 1889-1903.

Birk, S., W. Bonne, A. Borja, S. Brucet, A. Courrat, S. Poikane, A. Solimini, W. van de Bund, N. Zampoukas, D. Hering, 2012. Three hundred ways to assess Europe's surface waters: An almost complete overview of biological methods to implement the Water Framework Directive. Ecological Indicators, 18: 31-41.

Birk, S., D. Chapman, L. Carvalho, B. M. Spears, H. E. Andersen, C. Argillier, S. Auer, A. Baattrup-Pedersen, L. Banin, M. Beklioğlu, E. Bondar-Kunze, A. Borja, P. Branco, T. Bucak, A. D. Buijse, A. C. Cardoso, R.-M. Couture, F. Cremona, D. de Zwart, C. K. Feld, M. T. Ferreira, H. Feuchtmayr, M. O. Gessner, A. Gieswein, L. Globevnik, D. Graeber, W. Graf, C. Gutiérrez-Cánovas, J. Hanganu, U. Işkın, M. Järvinen, E. Jeppesen, N. Kotamäki, M. Kuijper, J. U. Lemm, S. Lu, A. L. Solheim, U. Mischke, S. J. Moe, P. Nõges, T. Nõges, S. J. Ormerod, Y. Panagopoulos, G. Phillips, L. Posthuma, S. Pouso, C. Prudhomme, K. Rankinen, J. J. Rasmussen, J. Richardson, A. Sagouis, J. M. Santos, R. B. Schäfer, R. Schinegger, S. Schmutz, S. C. Schneider, L. Schülting, P. Segurado, K. Stefanidis, B. Sures, S. J. Thackeray, J. Turunen, M. C. Uyarra, M. Venohr, P. C. von der Ohe, N. Willby, D. Hering, 2020. Impacts of multiple stressors on freshwater biota across spatial scales and ecosystems. Nature Ecology & Evolution, 4: 1060-1068.

Borja, A., 2018. Testing the efficiency of a bacterial community-based index (microgAMBI) to assess distinct impact sources in six locations around the world. Ecological Indicators, 85: 594-602.

Borja, A., J. Franco, V. Pérez, 2000. A marine biotic index to establish the ecological quality of soft-bottom benthos within European estuarine and coastal environments. Marine Pollution Bulletin, 40: 1100-1114.

Bourhane, Z., A. Lanzén, C. Cagnon, O. Ben Said, E. Mahmoudi, F. Coulon, E. Atai, A. Borja, C. Cravo-Laureau, R. Duran, 2022. Microbial diversity alteration reveals biomarkers of contamination in soil-river-lake continuum. Journal of Hazardous Materials, 421: 126789.

Clark, D. E., F. Stephenson, J. E. Hewitt, J. I. Ellis, A. Zaiko, A. Berthelsen, R. H. Bulmer, C. A. Pilditch, 2021. Influence of land-derived stressors and environmental variability on compositional turnover and diversity of estuarine benthic communities. Marine Ecology Progress Series, 666: 1-18.

Korpinen, S., L. Laamanen, L. Bergström, M. Nurmi, J. H. Andersen, J. Haapaniemi, E. T. Harvey, C. J. Murray, M. Peterlin, E. Kallenbach, K. Klančnik, U. Stein, L. Tunesi, D. Vaughan, J. Reker, 2021. Combined effects of human pressures on Europe's marine ecosystems. AMBIO, 50: 1325-1336.

Lanzén, A., I. Mendibil, Á. Borja, L. Alonso-Sáez, 2021. A microbial mandala for environmental monitoring: Predicting multiple impacts on estuarine prokaryote communities of the Bay of Biscay. Molecular Ecology, 30: 2969-2987.

Pearman, J. K., F. Afandi, P. Hong, S. Carvalho, 2018. Plankton community assessment in anthropogenic-impacted oligotrophic coastal regions. Environmental Science and Pollution Research, 25: 31017–31030.

Sun, J., X. Chen, J. Yu, Z. Chen, L. Liu, Y. Yue, Z. Fu, M. Yang, F. Wang, 2021. Deciphering Historical Water-Quality Changes Recorded in Sediments Using eDNA. Frontiers in Environmental Science, 9: 10.3389/fenvs.2021.669582.