Exploring and mapping plankton genomics data with Blue-Cloud

Pavla Debeljak, Alexandre Schickele, Sakina-Dorothée Ayata, Lucie Bittner, Jean-Olivier Irisson, Federico Drago

Blue-Cloud

Exploring and mapping plankton genomics data with Blue-Cloud

Authors:
Pavla Debeljak, Alexandre Schickele, Sakina-Dorothée Ayata, Lucie Bittner, Jean-Olivier Irisson, Sorbonne Université; Federico Drago, Trust-IT Services

Blue-Cloud has received funding from the European Union's Horizon programme call BG-07-2019-2020, topic: [A] 2019 - Blue Cloud services, Grant Agreement No.862409.

Recent metagenomic studies have revealed that marine plankton is far more diverse than previously thought (Carradec et al. 2018, Salazar et al. 2019, Duarte et al. 2020), with hundreds of thousands of genetically distinct taxa and more than 116 million genes documented for eukaryotic plankton and 47 million genes for prokaryotes. However, the taxonomy and/or function of more than half of the planktonic 'omic' sequences is still unknown. These unprecedented amounts of data on planktonic communities call for innovative, data-driven approaches to quantify and observe their biogeographic importance (Faure et al. 2021).

Marine plankton play a fundamental role in the global biogeochemical cycles and marine food webs. They are also a sentinel of environmental changes. Gathering more information about their genomics can help us better describe plankton distributions at global scale and further understand their response to environmental changes.

The Blue-Cloud demonstrator Plankton Genomics responds to this challenge by mining the rich metagenomic and metatranscriptomic data collected during the Tara Oceans mission and combining it with in situ or climatological environmental information to infer the function, taxonomy and distribution of the large portion of unknown sequences. In this article, we are going to explore the main results of the demonstrator and its intended evolution.

The demonstrator is led by the European Bioinformatics Institute (EMBL-EBI) and created by the Faculty of Sciences at Sorbonne University.

What can researchers do with it?

The Plankton Genomics demonstrator has developed two services within its dedicated Blue-Cloud Virtual Lab:

- Notebook 1: Exploring genetic data & identifying clusters containing unknown genes
  - Discovery of as yet undescribed biodiversity from genetic signals from the characterisation of their geographical distributions, co-occurrences/exclusions and correlation with environmental variables.
- Notebook 2: Mapping the geographic distribution of plankton functional gene clusters using habitat prediction models
  - Exploration of genetic markers of plankton diversity and abundance, in particular the new ones discovered above, to predict their spatial distribution and describe Essential Ocean Variables (EOVs) related to biological processes.

Notebook 1 - Exploring genetic data & identifying clusters containing genes of unknown functions

The key objective of this service is to enable the discovery of unknown genes using the large dataset collected during the Tara Oceans Expedition. The service allows retrieving genes of unknown functions from annotation files for 4 different plankton size classes and then building gene clusters by similarities, which correspond here to putative protein families present in the environments. These families can be further explored to tackle evolutionary or ecological questions, as well as to rebuild metabolic networks, which could serve as a base for biogeochemical cycle modelling.

The output file is then used for the second service of this demonstrator.

1.1) Retrieving Unknowns from annotation files

In Notebook 1.1, sequence functional annotations from the Marine Atlas of Tara Ocean Unigenes (MATOU) are used for the exploration of non-annotated sequences (Carradec et al. 2018). The sequences are available in FASTA format, which is a text-based format for representing either nucleotide sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes.

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The FASTA sequence files containing over 116 million genes are retrievable through the EMBL ENA web service.

The dataset includes 4 different size classes (Figure 1) based on the different filter sizes used at on-board filtration: 0.8-5 µm (pico- to nanoplankton), 5-20 µm (nanoplankton), 20-180 µm (microplankton), 180-2000 µm (mesoplankton). The Jupyter Notebook allows for the extraction of Unknowns based on Function or Taxonomy retrieved from Carradec et al.. The codes in the notebook find the unknown sequences and calculate the ratio of unknowns to knowns for different size fractions of the *Tara* Ocean data. Furthermore, giant scaffolds can be excluded and mean sequence length and standard deviation calculated and plotted in R (using the "ggplot" package).

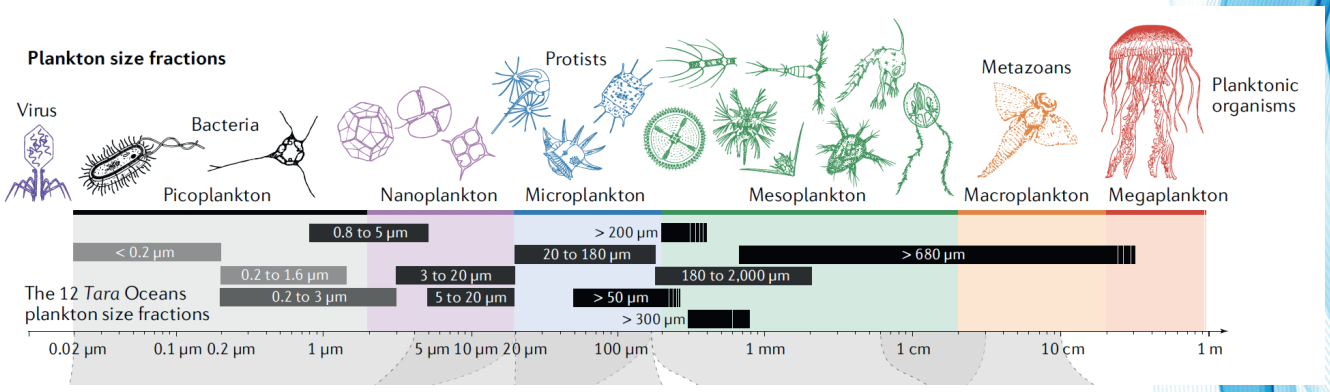

Figure 1. Size Fractions from Pico- to Megaplankton by filter size from the Tara Oceans expeditions. Picture modified from Sunagawa et al. 2020.

The different size classes (Fig.1) ranging from Nanoplankton to Mesoplankton analysed together or separately can then be plotted (Figure 2) using environmental data in R (leaflet package).

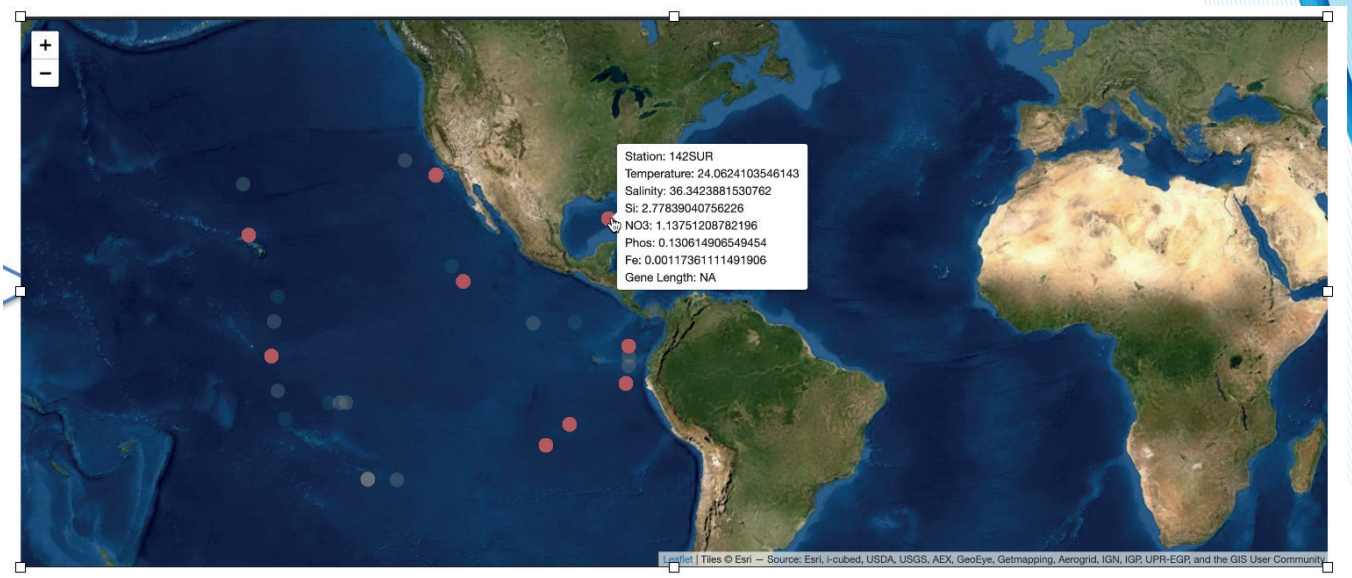

Figure 2. Example of the visualisation of Tara Oceans data and environmental parameters in the leaflet package in R implemented in Notebook 1.1.

1.2) Creating protein functional clusters

Notebook 1.2. allows for the creation of protein functional clusters from FASTA files derived from metagenomic and metatranscriptomic sequencing. These clusters contain annotated as well as unknown sequences that can be passed on to the IT service: mapping the geographic distribution of plankton functional gene clusters using habitat prediction models (described in the next section).

The necessary data can be retrieved from https://www.genoscope.cns.fr/tara under "Tara Oceans Eukaryotic Genomes ("MAGs") (Delmont et al. 2021 accessible through https://www.biorxiv.org/content/10.1101/2020.10.15.341214v2).

The provided peptide fast file contains 713 manually curated meta genome assembled genomes (or MAGs) containing 10,207.435 proteins. With these, over 10 million sequence functional clusters of proteins can be built using a sequence similarity network (here using the igraph package in R) (c.f. Methodology described in Faure et al. 2021).

In this graph representing the sequence similarity network (Figure 3, from Forster et al. 2015), nodes are protein sequences and edges represent the similarity and coverage between each pair of sequences.

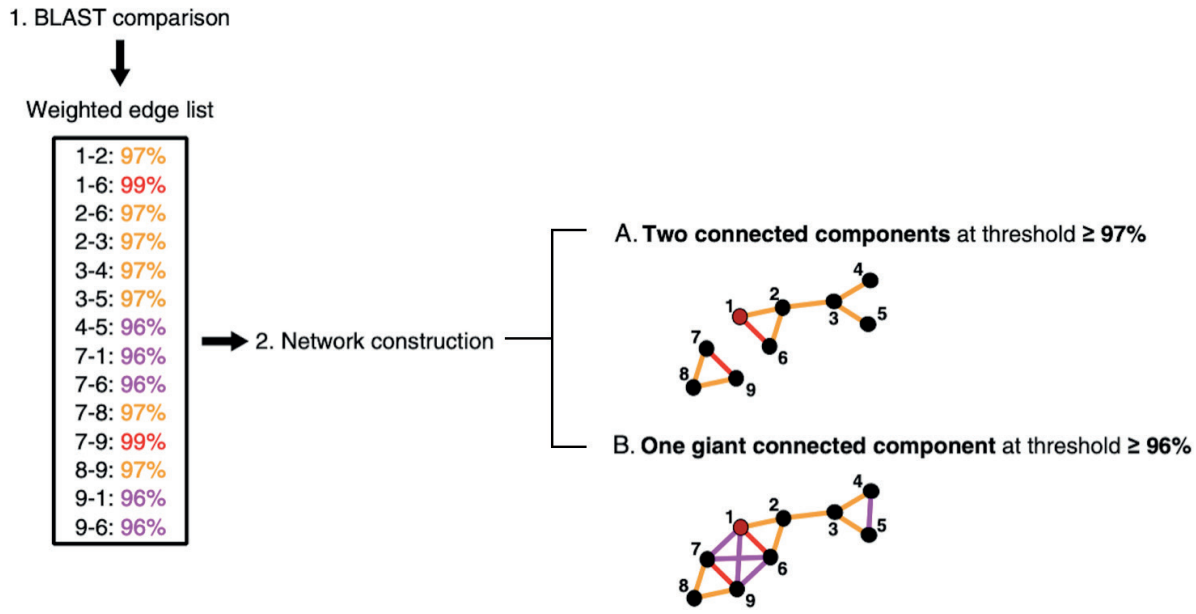

Figure 3 - Building a Sequence similarity network (picture modified from Forster et al. 2015). The mock dataset results in: A) two connected components when the minimum %ID required to connect two nodes is ≥97%; B) a single giant component when the minimum %ID required to connect two nodes is ≥96%

The sequence similarity networks display sequences as nodes (black nodes represent environmental sequences, red nodes represent sequences of cultured ciliates), connected by edges reflecting their %ID obtained from a BLAST analysis (see list of weighted edges in which the colour code reflects the %ID; red for 99%, orange for 97%, pink for 96%). The corresponding colour code is used on the networks (right panel of the figure) to explore and structure the data.

Such approaches allow for the observation of sequence clusters that are putatively homogenous in function and have recently highlighted the potential for deciphering global ocean biogeochemistry (Faure et al. 2021).

Notebook 2 - Mapping the geographical distribution of plankton functional gene clusters using habitat models

The concentrations of planktonic organisms and, by extension, the genetic composition of planktonic assemblages in a given geographical area is influenced by its environmental context. This notebook provides tools to define the relationships between the abundance of plankton genes and environmental variables, in order to then project their potential biogeography. It starts by focusing on a key metabolic pathway (e.g. Oxidative phosphorylation, Carbon fixation) and then informs about as-yet-unknown plankton genes possibly related to this pathway.

The relationship between the abundance of plankton gene clusters and environmental variables is estimated through a machine learning regression method, namely multivariate gradient boosting (i.e. several gene clusters related to the same metabolic pathway are modelled at once; Nespoli & Medici, 2020), and is implemented in R, on top of a Python library (MBTR).

Following best practices in machine learning, we automatically select the best set of hyperparameters for the model through grid-search and *k*-fold cross-validation. A grid of potentially suitable algorithm hyperparameters is defined; for each combination of hyperparameters the model is evaluated using cross-validation: the data

are split in *k* equal sized groups, *k*models are trained each on *k*-1 splits, holding the last split for evaluation. The quality of fit of each model is measured by a loss function at each boosting round and the best hyperparameters combination is selected as the one that produced the minimum loss, averaged over all cross-validation folds.

The final performance of the selected model is quantified on the evaluation split of each cross-validation fold and the following metrics: the multivariate R-squared (R2) and Root Mean squared Error (RMSE).

Once this best model is selected and deemed of high enough quality, it is used to predict the potential distribution of all gene clusters modelled over the entire ocean, based on the application of the fitted relationship with environmental variables. Bootstraps of the original data are used to estimate variability in the prediction according to slight variations in the input data.

As output, this notebook produces a series of maps corresponding to the spatial distribution (average and uncertainty) of each selected gene cluster and their related functional description (**Figure 4**). To complement the functional interpretation of these outputs, additional descriptions of the selected gene clusters are available in the data, such as taxonomic classification and supplementary KEGG annotations (https://www.genome.jp/kegg/).

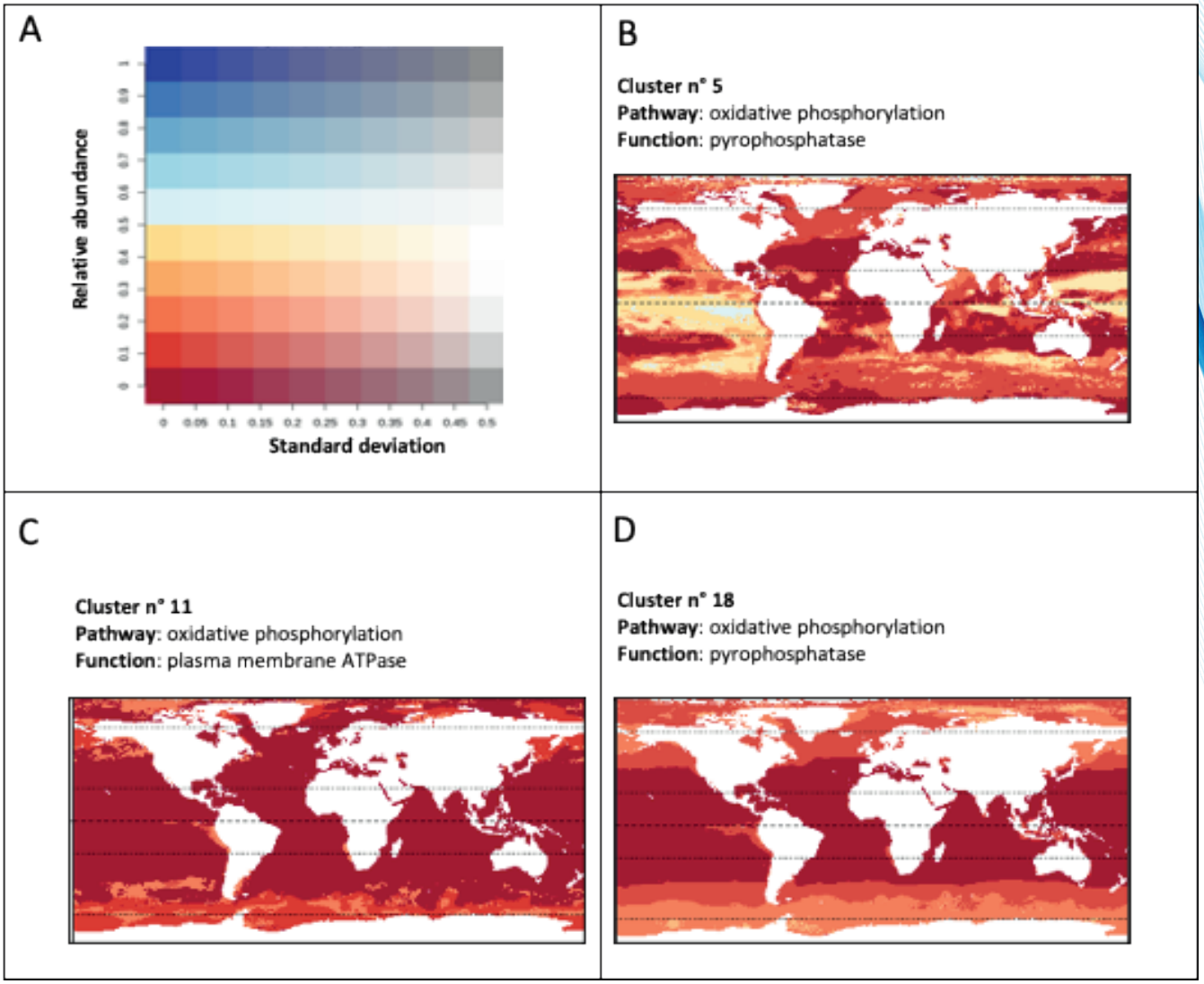

Figure 4: *Synthetic example outputs of notebook 2 with* ***(a)*** *the map legend and* ***(b-d)*** *the modelled distribution of the most abundant clusters of genes and their functional description, among a set of 19 gene clusters related to oxidative phosphorylation.*

Open for testing

The services in this Virtual Lab are mainly designed for expert users with a strong genomic, bioinformatics, and modelling/machine learning background, not because they are particularly difficult to use but because their *interpretation* requires such knowledge.

Still, the end-users include a broad base of scientists in quest of the identification of unknown sequences in the oceanic environment, and also interested, for example in plankton biogeography, marine biogeochemistry, ecosystem health, and climate science.

The Virtual Lab is now open for testing and accessible via the Blue-Cloud Virtual Research Environment on D4Science.

🔗 Test the Plankton Genomics Virtual Lab

References

1. Carradec, Q., Pelletier, E., Da Silva, C. et al. (2018) A global ocean atlas of eukaryotic genes. Nat Commun 9, 373. https://doi.org/10.1038/s41467-017-02342-1
2. Duarte, C.M. et al. (2020) Sequencing effort dictates gene discovery in marine microbial metagenomes. SFAM 22:11 https://doi.org/10.1111/1462-2920.15182
3. Faure, E., Ayata, SD. & Bittner, L. (2021). Towards omics-based predictions of planktonic functional composition from environmental data. Nat Commun 12, 4361. https://doi.org/10.1038/s41467-021-24547-1
4. Forster D., Bittner L., Karkar S., Dunthorn M. , Romac S., Audic S., Lopez P., Stoeck T. & Bapteste E. (2015). Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. BMC Biology 13:16. https://doi.org/10.1186/s12915-015-0125-5
5. Sunagawa, S., Acinas, S.G., Bork, P. et al. (2020) Tara Oceans: towards global ocean ecosystems biology. Nat Rev Microbiol 18, 428–445. https://doi.org/10.1038/s41579-020-0364-5
6. Nespoli, L. and Medici, V. (2020). Multivariate Boosted Trees and Applications to Forecasting and Control . arXiv preprint arXiv:2003.03835