Hackathon - Dataset description

Here is a technical description of the provided dataset. Understanding the technical terms here are not necessary to take part in the competition. When looking at your results, it might be relevant to take a second look at this file in order to give more relevant comments and explanations on your findings.

Data

There are 2 files:

- `abundance_table.csv` with the abundance for each organism present in each sample. In other words, it is the percentage of DNA in the sample that belongs to a given organism.
    - Column 0 `Scientific Name` is the name of the organism studied. E.g. "Homo sapiens".
    - Column 1 `Taxonomic Rank` is the taxonomic rank. The letter corresponds to (from the most general to most specific) "Domain, Kingdom, Phylum, Class, Order, Family, Genus, Species" and the number corresponds to the subhierarchy within the rank. E.g. "S4"
    - *Each one of the remaining columns* corresponds to a sample. The values taken in these columns refer to the percentage of this *organism* (a given row) in the composition of the *sample* (a given column).
        - The samples come from different aquaculture systems represented in the sample name by the first letter "A" or "B", for different treatment represented after the letter by "1", "2" or "3" and at different sampling time "_1", "_2", "_3" and in triplicate "A", "B" or "C".
        - For example, if a sample is called "A1_1A" it means that it is from system A, treatment 1, timepoint 1 and first replicate A. Wikipedia: "*In biology, a replicate is an exact copy of a sample that is being analyzed, such as a cell, organism or molecule, on which exactly the same procedure is done. This is often done in order to check for experimental or procedural error.*"
- `env_parameter_sample.csv` with the metadata (time, temperature, pH, salinity, etc.) corresponding to the conditions relative to each sample (see "Column 19 `sample_name`"). Since the triplicate sample are taken at the same time, the conditions of sampling were the same, so the sampling condition for "A1_1A", "A1_1B" and "A1_1C" would be denoted as "A1_1" for example (last column of the file).
    - Column 0 `time`: sampling date.
    - Column 1 `TAN_DF`: Total Ammonia Nitrogen in the drumfilter in mg/l.
        - "TAN" stands for "Total Ammonia Nitrogen" it is how much NH4+ and NH3 there is in the system at the time of the sampling. In recirculating aquaculture system, TAN is closely monitored because ammonia can be toxic for the fish. TAN is removed by nitrifying bacteria (so some of the organisms that we would expect to find in our sample).
        - "DF" stands for drum filter which is the compartment located before the biofilter (where the nitrifiers are).
    - Column 2 `pH_DF`: pH in the drumfilter.
    - Column 3 `NO2_DF`: NO2 in the drumfilter, in mg/l. The nitrifiers break down ammonia in a two-step process, first ammonia is oxidized in nitrite (NO2) by the ammonia oxidizing bacteria and then nitrite is oxidized into nitrate (NO3) by the nitrite oxidizing bacteria. NO2 is also toxic so it

    is also closely monitored. High nitrite in an aquaculture system indicates that the biofilter is not working properly.
    - Column 4 `Alkalinity_DF`: Alkalinity measured in "mg/l as CaCO3" in the drumfilter. Within the experiment we were trying to run different alkalinity level 70, 100, 200 as we wanted to see what would be the impact of different alkalinity level on the microbial community composition.
    - Column 5 `TAN_DGS`: Total Ammonia Nitrogen measured after the biofilter in mg/l.
    - Column 6 `pH_DGS`: pH measured after the biofilter.
    - Column 7 `NO2_DGS`: NO2 measured after the biofilter, in mg/l.
    - Column 8 `Alkalinity_DGS`: alkalinity in "mg/l as CaCO3" after the biofilter.
    - Column 9 `Flow_rate`: flow rate in L/min.
    - Column 10 `Treatment`: corresponds to the treatment we were targeting (70,100,200), which should be the same as the `Flow_rate` in theory, but of course differs in practice.
    - Column 11 `Module`: corresponds to the system from which the biocarriers were taken.
    - Column 12 `TAN_removal_biocarrier`: is the removal rate of TAN per biocarriers which are plastic beads where the biofilm attached in which the nitrifiers grow, so it is what we have sampled.
    - Column 13 `co2_mgl`: CO2 in mg/L before the biofilter.
    - Column 14 `h2s_ugl`: H2S in ug/L before the biofilter. A very toxic gas.
    - Column 15 `o2_mgl`: O2 in mg/L before the biofilter.
    - Column 16 `o2_sat`: O2 saturated in % before the biofilter.
    - Column 17 `salinity`: salinity before the biofilter.
    - Column 18 `temp`: temperature before the biofilter, in °C.
    - Column 19 `sample_name`: sample name (see other file `abundance_table.csv`).

Remarks

- The sample "AO" is a sort of reference sample, an "empty" sample.
- In the `abundance_table.csv` file, "NA" values actually mean "0", i.e. no trace of DNA from that given organism was found in this sample.
- Some samples have "R" at the end of their name (e.g. "B9_3BR"), it means that it is another replicate (here a replicate of "B9_3B").