MARCO-BOLO Data Analysis Challenge

MARCO-BOLO Data Analysis Challenge Overview

- **Goal**: MARCO-BOLO is a project that helps scientists understand what lives in the ocean by studying **environmental DNA** (eDNA), which is the DNA left behind by plants and animals in the water.

- **Challenge**: We want to develop a tool/workflow to validate the biodiversity of different marine bodies of water using eDNA data sets.

What is eDNA?

- **eDNA**: eDNA is DNA that comes from animals and plants in their environment (like water or soil). By studying this DNA, we can figure out what kinds of species are living in that area without having to see or catch them. (really helpful for small marine organisms)

- **Why eDNA Matters**: eDNA makes it easier to track what species live in the ocean. It's cheaper, faster, and less harmful than traditional methods like catching animals.

How Do We Analyze eDNA Data?

Data Example: We start with a sample from the ocean. It’s full of DNA sequences from different organisms. Here's what a small part of it looks like:

AGCTGATCG...

ATCGTACGT...

Our Job: We will use a simple program to group these sequences by species and figure out which animals or plants they belong to. For example:

- Sequence 1: Fish species A
- Sequence 2: Seaweed species B

Tools You’ll Use: Python, QIIME2, DADA2, etc

Workflow

What is the Workflow?

1. **Step 1: Collect eDNA** – We get water samples from the ocean.
2. **Step 2: Extract DNA** – We remove the DNA from the water and sequence it (this gives us long strings of letters like A, T, C, and G).
3. **Step 3: Analyze the Data** – We use tools to figure out which species the DNA comes from.

Data Files We'll Work With:

- **FASTA**: This is a file format that stores DNA sequences. Each sequence represents the DNA of a species. It looks like this:

```
>Sequence_1

AGCTGATCGTACG...

>Sequence_2

TACGCGTATGCTAG...
```

Tools We will Use

- **Tools for the Challenge**:
  - **QIIME 2**: A tool that cleans up the DNA data and groups similar sequences together.
  - **DADA2**: This tool finds the exact DNA sequences and helps us identify which species they belong to.
  - **MetaPhlAn**: A program that can quickly tell us which microbes (tiny organisms) are in the sample.
- **File Formats**:
  - **FASTA**: Stores the DNA sequences.
  - **OTU/ASV Tables**: Shows how many times each DNA sequence appears and what species it belongs to.
  - **Taxonomic Info**: A table that matches sequences to species names.
- **Resources We Use**:
  - **Databases**: We use large collections of DNA sequences from species (like the **SILVA** or **NCBI** databases) to match the DNA we found in the water to known species.
  - **Computers**: We run these tools on powerful computers to handle large amounts of DNA data.

OTU (Operational Taxonomic Unit): This groups similar DNA sequences to represent species. OTUs help us count how many different species are in the sample.

ASV (Amplicon Sequence Variant): ASVs are more exact than OTUs. They identify each unique DNA sequence, even if the difference is tiny, like one letter.

Questions?