Good and Bad Data Practices

Discussion and Activity

What is “bad” data?

Bad data

- Incorrect or incomplete data
  - Data are missing or wrong
- Improperly formatted data
  - Bespoke file formats, errors in file formatting, etc.
- Metadata do not properly describe data
  - Cannot map metadata to data
  - Metadata cannot differentiate between samples
- Data released without a use license
  - Unclear to potential users what can be done with the data

Bad data is not:

- Data that do not support the main hypothesis
- Data that are collected but not used in a study

What are some causes of bad data?

Potential causes of bad data

- Inadequate or inconsistent record keeping
- Data entry errors
- Incorrect data conversions/transformations
    - Can be file type -> file type or measurement -> measurement
- Data were collected using improper methods or bad tools
    - Expired reagents
    - Incorrect pH
- Equipment failure, limited supplies, etc.
- Incomplete knowledge of data licenses

Spot the difference

Laboratory Example 1

You are reporting methods for testing the pH sensitivity of an enzyme
(V-BrPO: Vanadium-dependent bromoperoxidase)

The pH sensitivity of C. officinalis V-BrPO activity was determined by altering the pH of the 50 mmol l-1 MES buffer containing 0.5 mU ml-1 of V-BrPO to between pH 5.8 and 7.8 at 25°C.

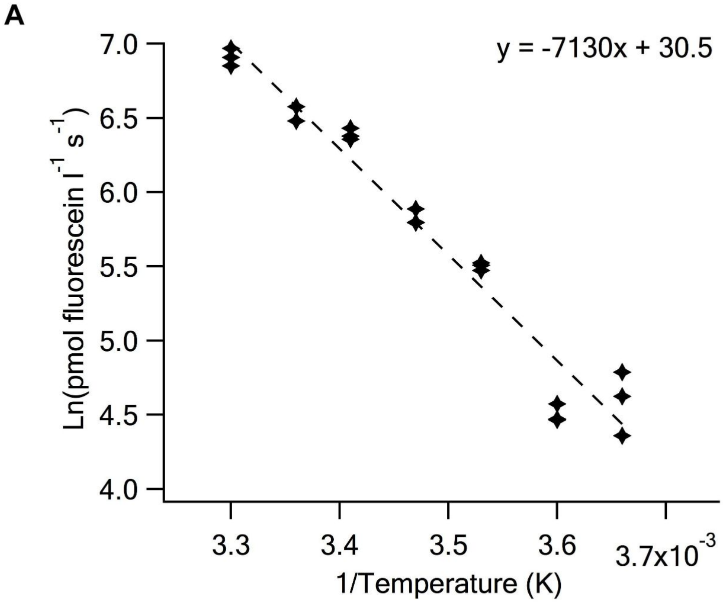
The pH sensitivity of C. officinalis V-BrPO activity was determined by altering the pH of the 50 mmol l-1 MES buffer containing 0.5 mU ml-1 of V-BrPO to between pH 5.8 and 7.8 at 20°C.

https://www.frontiersin.org/articles/10.3389/fmars.2019.00068/full

Temperature affects measured pH values and enzyme activity rates.

Temperature (ºC)	Measured pH of Water
0	7.47
25	7.00
50	6.63
100	6.14

https://chem.libretexts.org/



Laboratory Example 2

You are reporting growing conditions for two polar diatom strains.

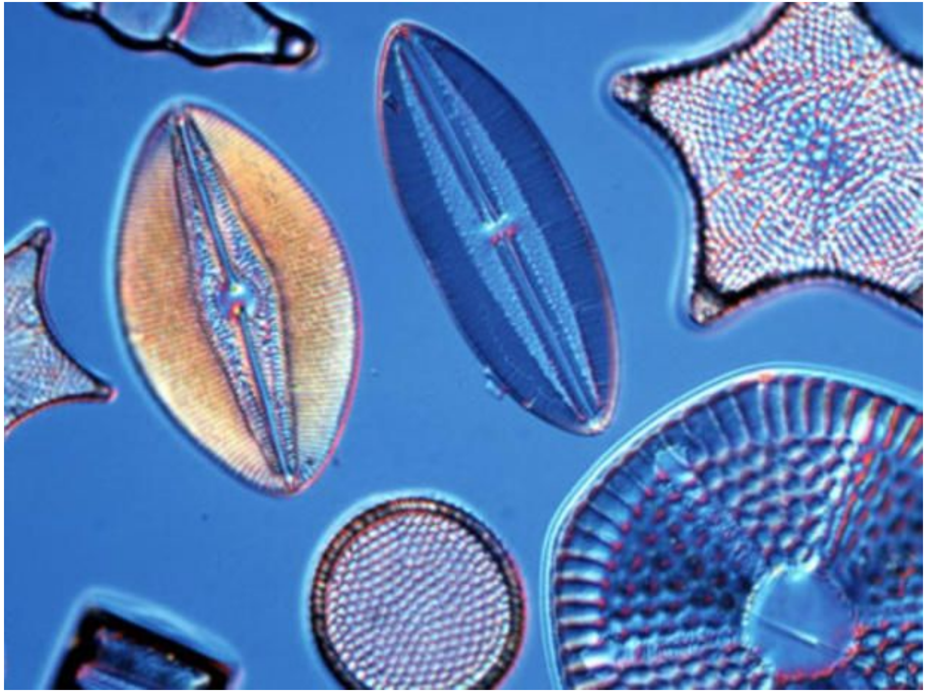
Diatom strains were grown in L1 medium, at 4°C under a 14h:10h light:dark cycle and light intensity of 60 μmol photons m-2 s-1.

Diatom strains were grown in LB medium, at 4°C under a 14h:10h light:dark cycle and light intensity of 60 μmol photons m-2 s-1.

https://www.frontiersin.org/articles/10.3389/fmars.2019.00068/full

L1 is an enriched seawater medium – perfect for diatoms, which are marine microbes.

LB is a nutrient rich medium commonly used to grow bacteria.



Laboratory Example 3

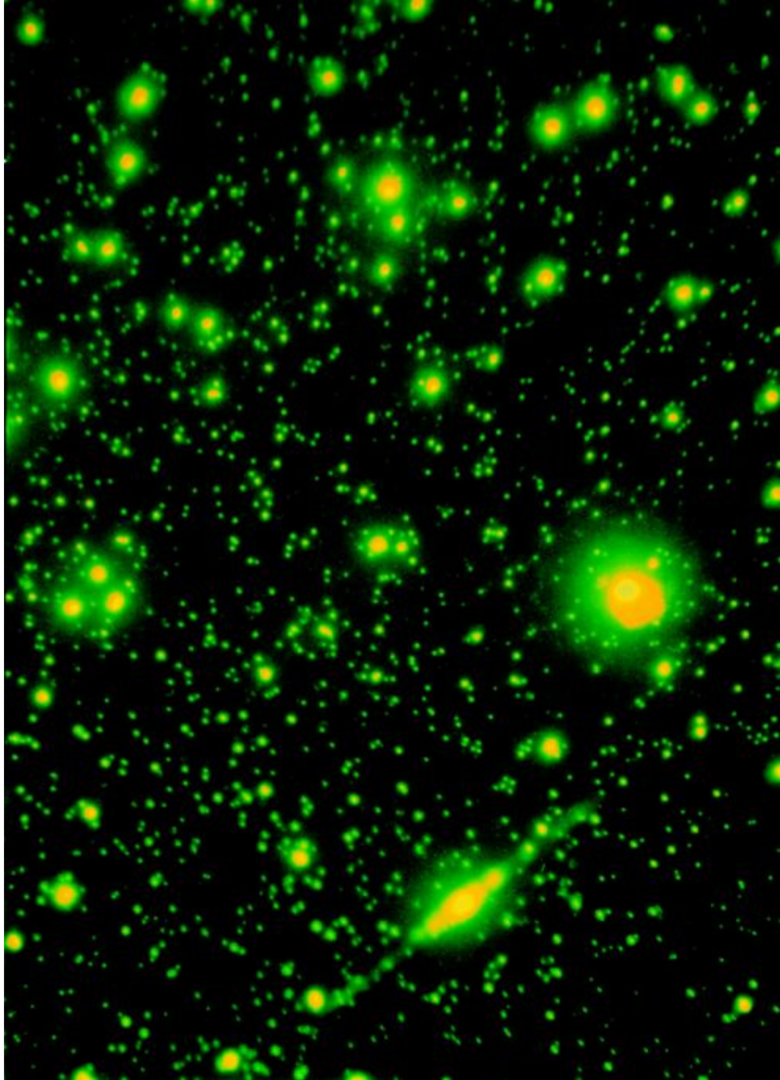
You are reporting on the filtration steps for removing microbes from your sample to make a virus-only sample.

10 liters of freshwater from an agricultural pond were sampled monthly, and filtered sequentially through 1 and 0.02 µm filter membranes.

10 liters of freshwater from an agricultural pond were sampled monthly, and filtered sequentially through 1 and 0.22 µm filter membranes.

https://www.frontiersin.org/articles/10.3389/fmicb.2018.00792/full

Viruses will also get stuck in the 0.02 µm filter, so you will create virus-free seawater instead of virus-only seawater.



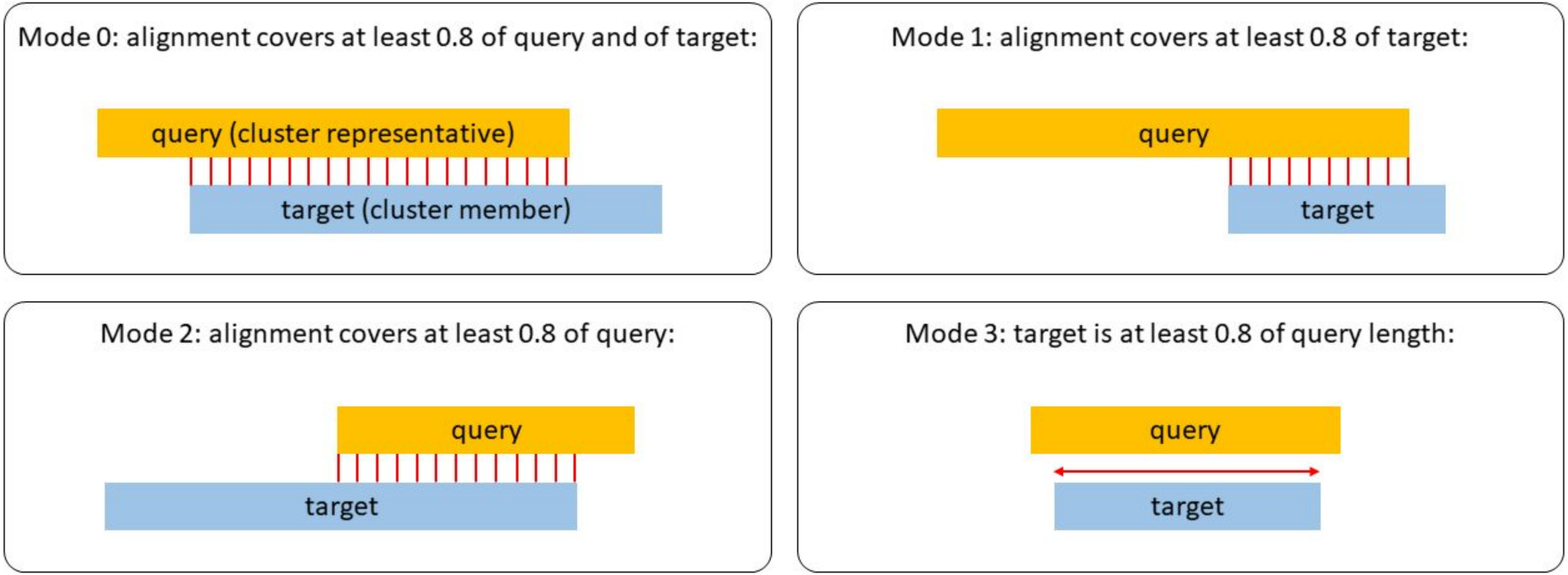
Bioinformatics Example

You are reporting the command you used to cluster protein sequences.

```
mmseqs easy-cluster examples/DB.fasta clusterRes tmp
--min-seq-id 0.5 -c 0.8 --cov-mode 1
```

```
mmseqs easy-cluster examples/DB.fasta clusterRes tmp
--min-seq-id 0.5 -c 0.8 --cov-mode 3
```

Different coverage modes lead to different clustering results



Mode 0: alignment covers at least 0.8 of query and of target:

Mode 1: alignment covers at least 0.8 of target:

Mode 2: alignment covers at least 0.8 of query:

Mode 3: target is at least 0.8 of query length:

https://github.com/soedinglab/mmseqs2/wiki#clustering-format

Good computer practices

Computational courtesy

- Do not use spaces in filenames
    - Use _underscores_ instead
- Use normal file extensions
    - Do not make up your own extensions
- Do not open text documents in word processors (e.g., Word, Docs)
    - Word processors add whitespace and other hidden characters
- The computer is not wrong
    - Though you may very occasionally find a software bug