Data Accessibility Status

Joanna Colovas

November 21st 2023

Project Goals

- Use mikropml (ML) R package to classify ASM papers as "containing new sequence data" or not based on most common "grams" (words) from the paper
- Report statistics on number of citations/paper as a function of data availability

Current Status

- Still need to be done from Adena (20230731)
  - Complete testing of Random Forest(RF) model on classification of data availability (with tuning of hyperparameters)
  - Need to classify more papers to investigate false positives/false negatives from model, and appearances of duplicate papers

Proposed Methodology

- Scrape ASM papers from Web of Science
- Sort alphabetically by title to remove duplicates
- Create list of 'grams' for each paper
  - See if there's a way to keep common words/phrases together based on HTML (i.e *Vibrio cholerae*, Illumina MiSeq)
- Use mikropml to create a model of the most predictive words that aid in classifying papers **which model?**
  - Ensure feature table is generated, and comb the top X words by hand (that give model X% sensitivity/specificity)
  - Better to acknowledge that it's not going to be perfect but disclose error rate of Y% (10% error? 5%? Let model tell us?)

Proposed Methodology

- Use generated lists from model to write script that will classify the rest of the papers from ASM
- Statistical analysis after paper classification, look at relationship between data availability and avg number of citations
  - Does this vary by ASM journal?