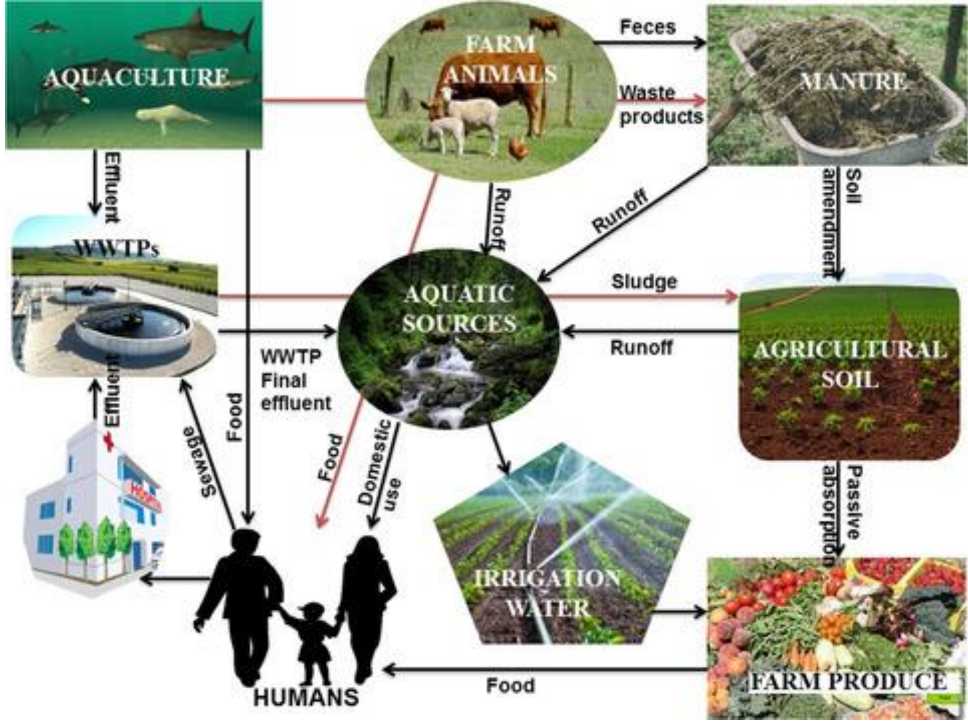
Predicting ARG composition in Effluent based on Influents in WWTP

Monjura Afrin Rumi

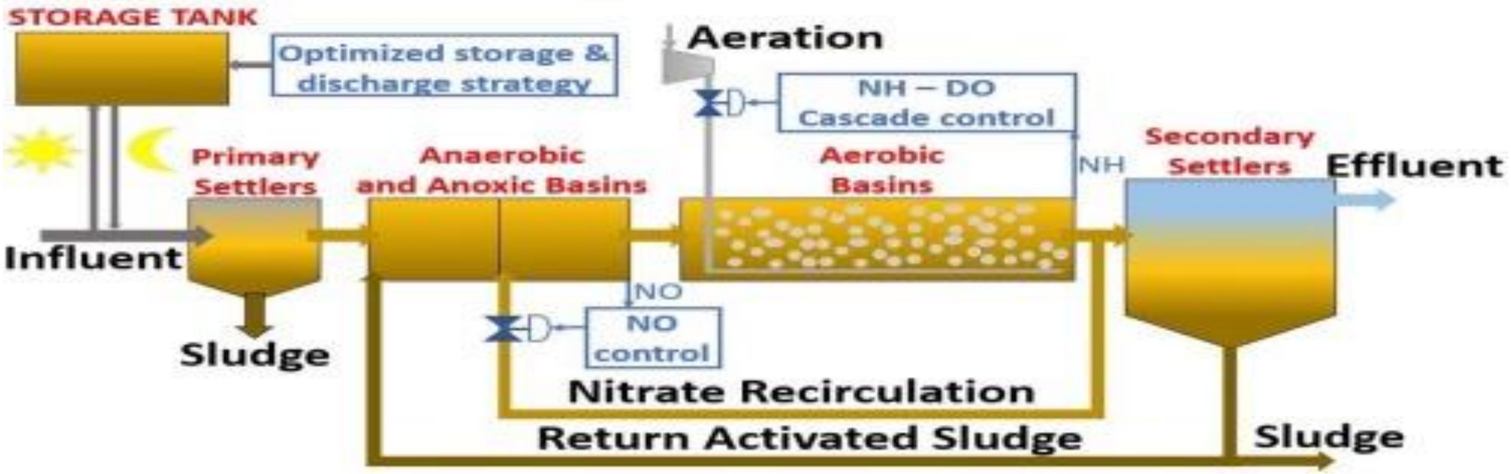
Introduction

- ARG – antibiotic resistance gene
- WWTP – wastewater treatment plant
- Influent - represents ARGs carried population
- Effluent - discharged into the environment

1



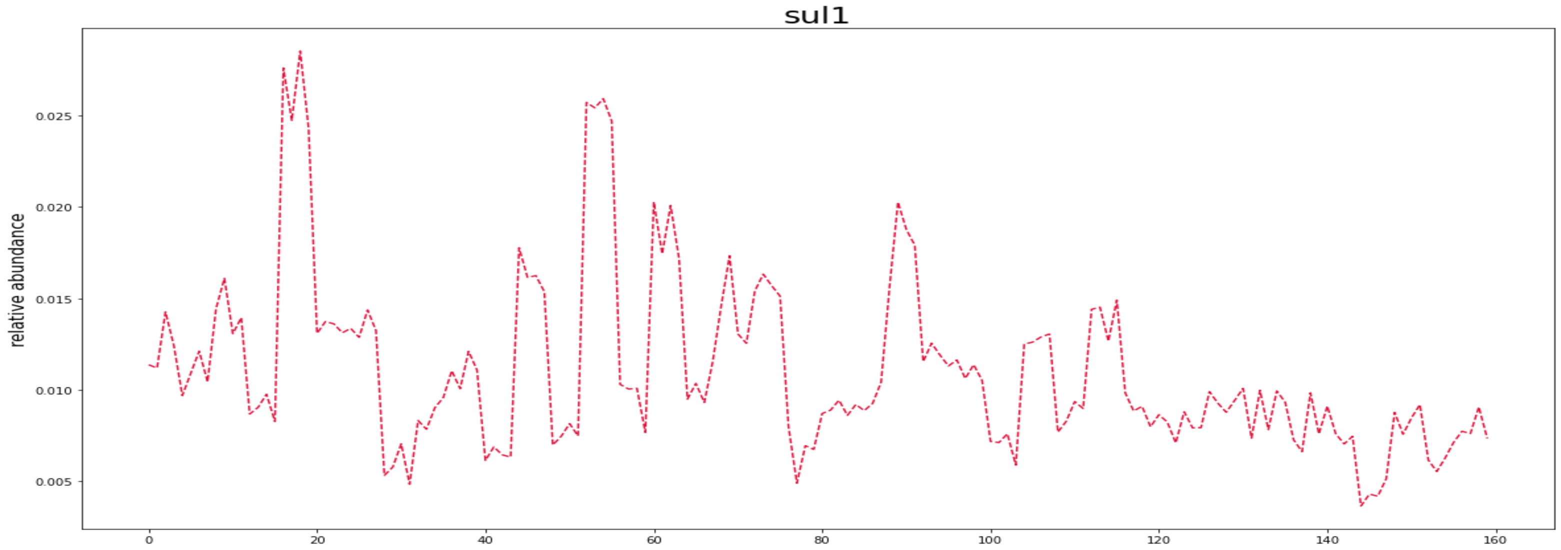
2



1. Chidozie D. Iwu,Lise Korsten,Anthony I. Okoh. The incidence of antibiotic resistance within and beyond the agricultural ecosystem: A concern for public health. Microbiology, Volume9, Issue9, September 2020.
2. Melinda Simon-Várhelyi, Vasile Mircea Cristea, Alexandra Veronica Luca. Reducing energy costs of the wastewater treatment plant by improved scheduling of the periodic influent load. Journal of Environmental Management, Volume 262,2020.

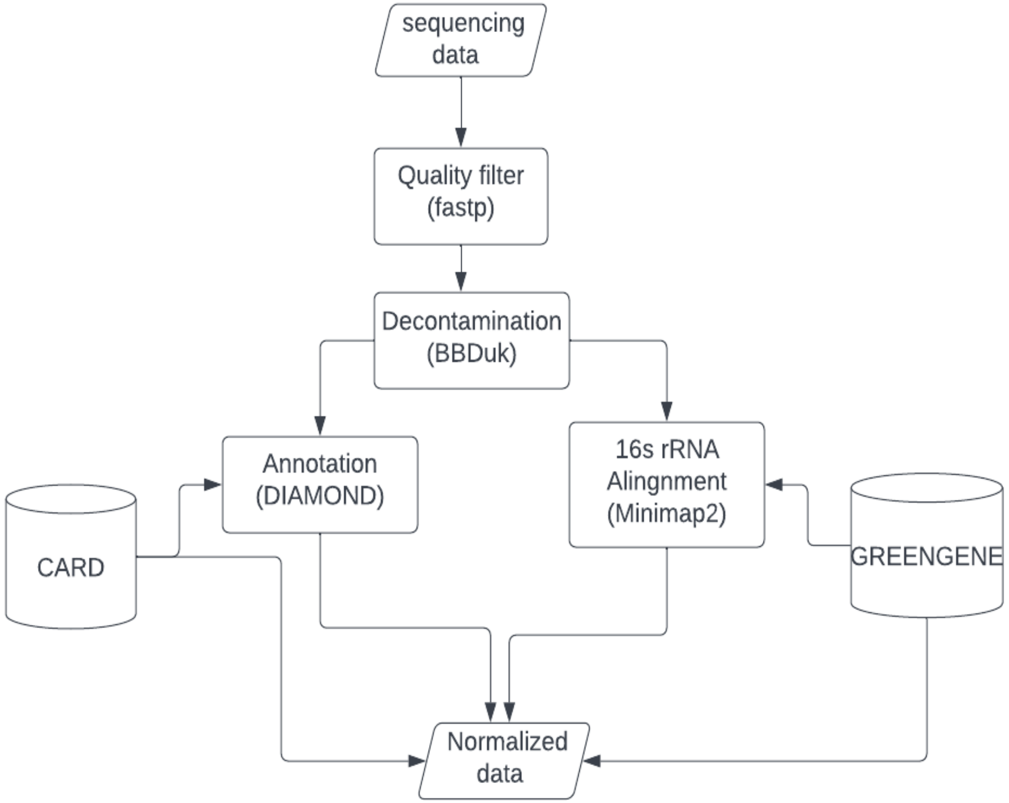
Introduction

- Target: predict ARG abundances in effluent from influent
- Limitations
  - Sequencing is expensive
  - Not enough data
  - Data aggregation is non-trivial



Data

- Data Collection
  - Christianburg, VA
  - August, 2020 – September, 2021
  - 192 effluent samples
  - 224 influent samples



sequencing data

Quality filter (fastp)

Decontamination (BBDuk)

Annotation (DIAMOND)

16s rRNA Alingnment (Minimap2)

CARD

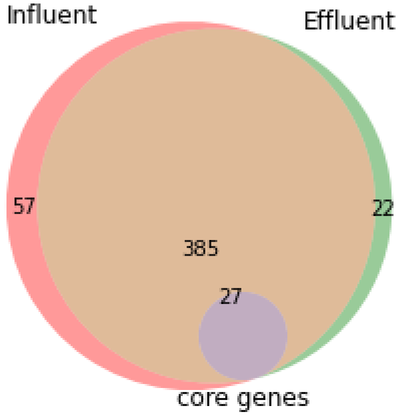
GREENGENE

Normalized data

$$\text{Abundance}=\sum_{1}^{n}\frac{N_{\text{ARG-like sequence}}\times L_{\text{reads}}/L_{\text{ARG reference sequence}}}{N_{\text{16S sequence}}\times L_{\text{reads}}/L_{\text{16S sequence}}}$$

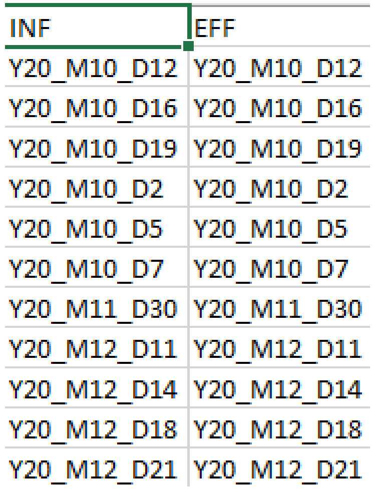
Preprocessing

- Filtering
  - Genes with low count
  - threshold = 5; Rarefaction [R, vegan]
  - ~500 unique genes available
- Feature and target selection
  - Use of domain knowledge
    - Core genes – present in all samples
    - Mutually exclusive gene
    - Low frequency – present in a few samples
  - Use ML technique
    - Correlation, mutual information

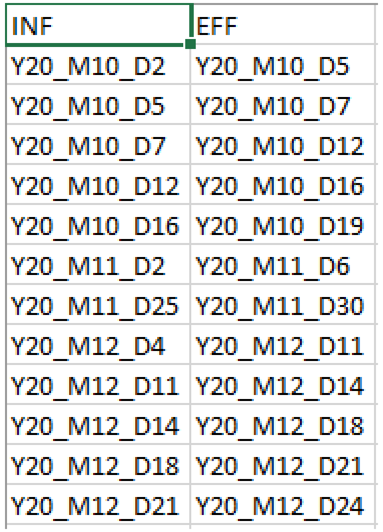


Preprocessing

- Metadata
  - Time dependency
  - Impacts X, y formation
    - Same day - 160 samples
    - Gap day - 156 samples
      - 2-5 days gap



INF	EFF
Y20_M10_D12	Y20_M10_D12
Y20_M10_D16	Y20_M10_D16
Y20_M10_D19	Y20_M10_D19
Y20_M10_D2	Y20_M10_D2
Y20_M10_D5	Y20_M10_D5
Y20_M10_D7	Y20_M10_D7
Y20_M11_D30	Y20_M11_D30
Y20_M12_D11	Y20_M12_D11
Y20_M12_D14	Y20_M12_D14
Y20_M12_D18	Y20_M12_D18
Y20_M12_D21	Y20_M12_D21



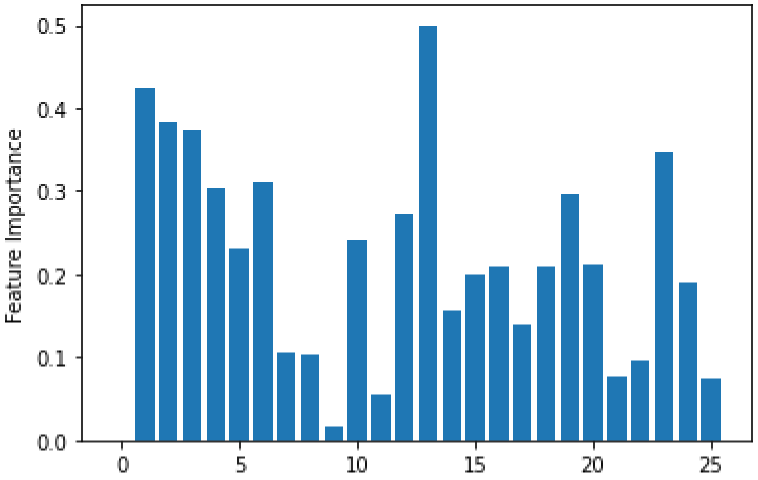
INF	EFF
Y20_M10_D2	Y20_M10_D5
Y20_M10_D5	Y20_M10_D7
Y20_M10_D7	Y20_M10_D12
Y20_M10_D12	Y20_M10_D16
Y20_M10_D16	Y20_M10_D19
Y20_M11_D2	Y20_M11_D6
Y20_M11_D25	Y20_M11_D30
Y20_M12_D4	Y20_M12_D11
Y20_M12_D11	Y20_M12_D14
Y20_M12_D14	Y20_M12_D18
Y20_M12_D18	Y20_M12_D21
Y20_M12_D21	Y20_M12_D24

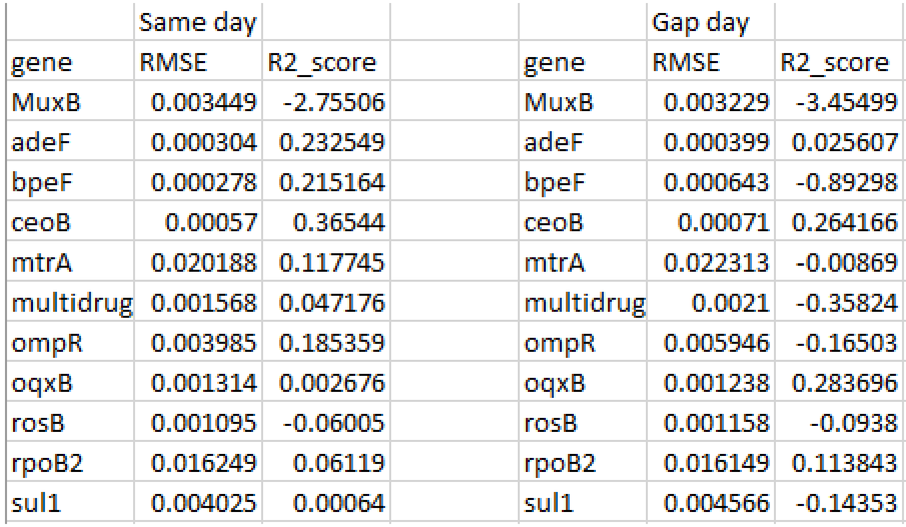
Method

- Alignment of input-output
  - Algorithm: Linear Regression
  - X & y formation
    - Feature: one/multiple INF core genes
    - Target: corresponding EFF core gene
  - Evaluation metric
    - RMSE
    - R-square

Result: Linear Regression

- Feature/target 1:1
  - Negative R-square
- Feature/target 10:1
  - k best feature; mutual information





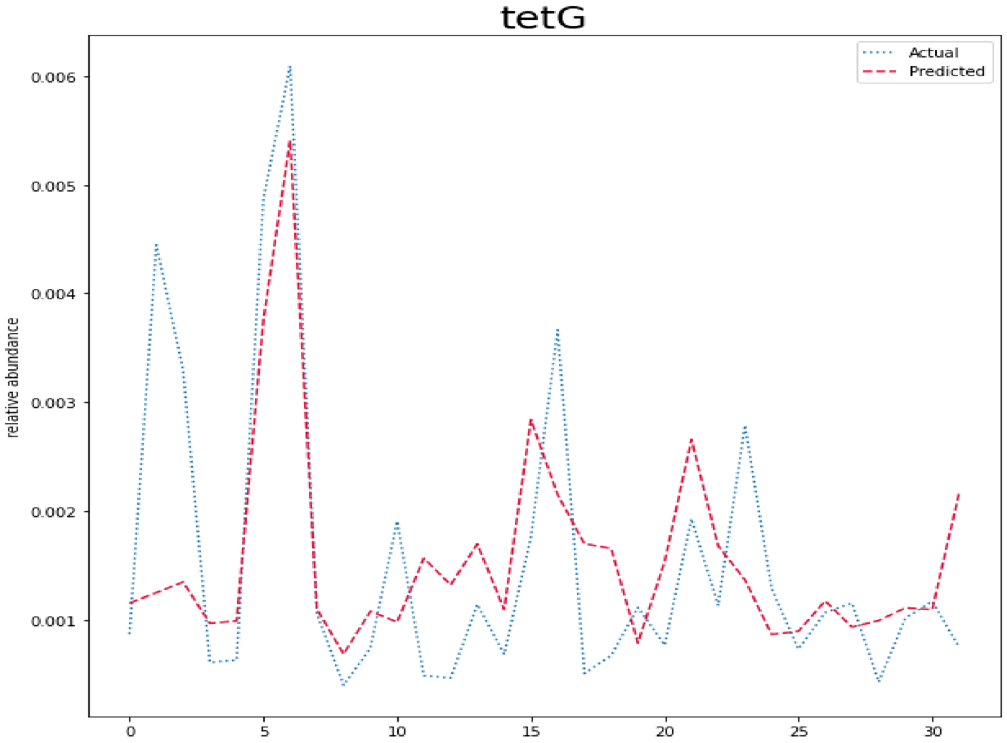
	Same day				Gap day	
gene	RMSE	R2_score		gene	RMSE	R2_score
MuxB	0.003449	-2.75506		MuxB	0.003229	-3.45499
adeF	0.000304	0.232549		adeF	0.000399	0.025607
bpeF	0.000278	0.215164		bpeF	0.000643	-0.89298
ceoB	0.00057	0.36544		ceoB	0.00071	0.264166
mtrA	0.020188	0.117745		mtrA	0.022313	-0.00869
multidrug	0.001568	0.047176		multidrug	0.0021	-0.35824
ompR	0.003985	0.185359		ompR	0.005946	-0.16503
oqxB	0.001314	0.002676		oqxB	0.001238	0.283696
rosB	0.001095	-0.06005		rosB	0.001158	-0.0938
rpoB2	0.016249	0.06119		rpoB2	0.016149	0.113843
sul1	0.004025	0.00064		sul1	0.004566	-0.14353

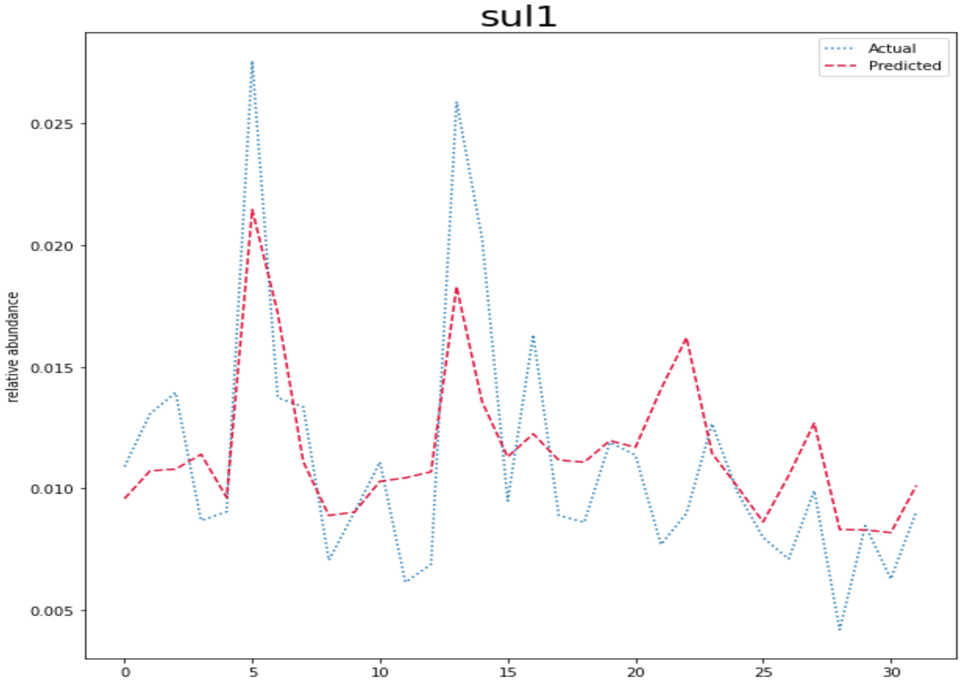
Method (continued)

- Algorithm
  - Random forest
  - MultiOutputRegressor (scikit-learn)
- X & y formation
  - highly abundant genes
  - Threshold = 50%, Grid search
  - Common in INF & EFF
  - 93 genes available

Result: Random forest

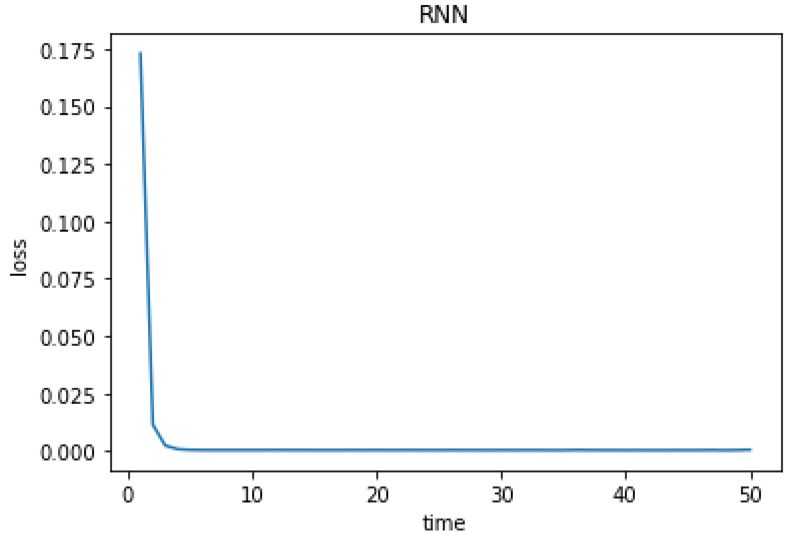
- Evaluation metric
  - RMSE: 0.0028
  - R-square: 0.75



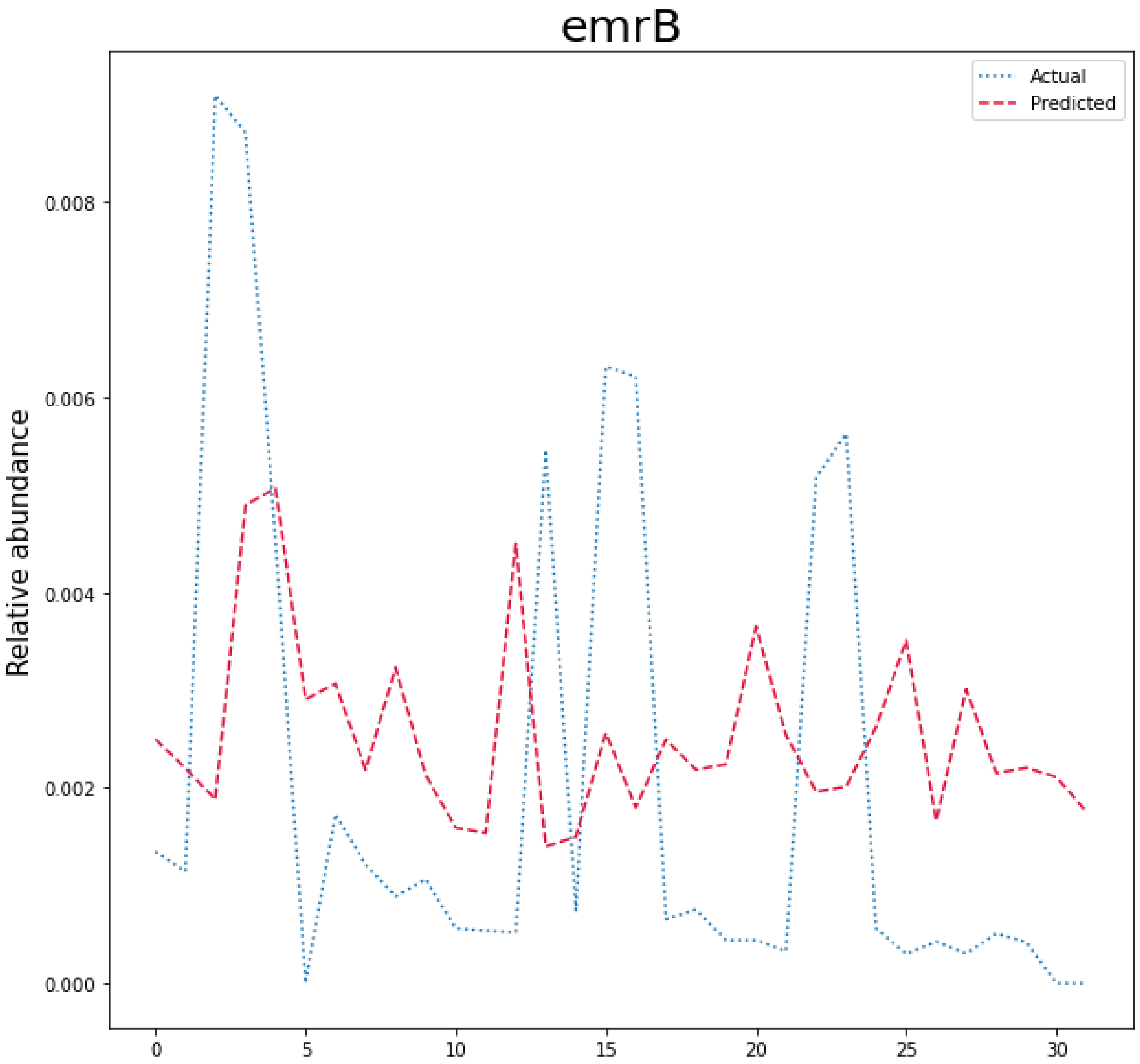


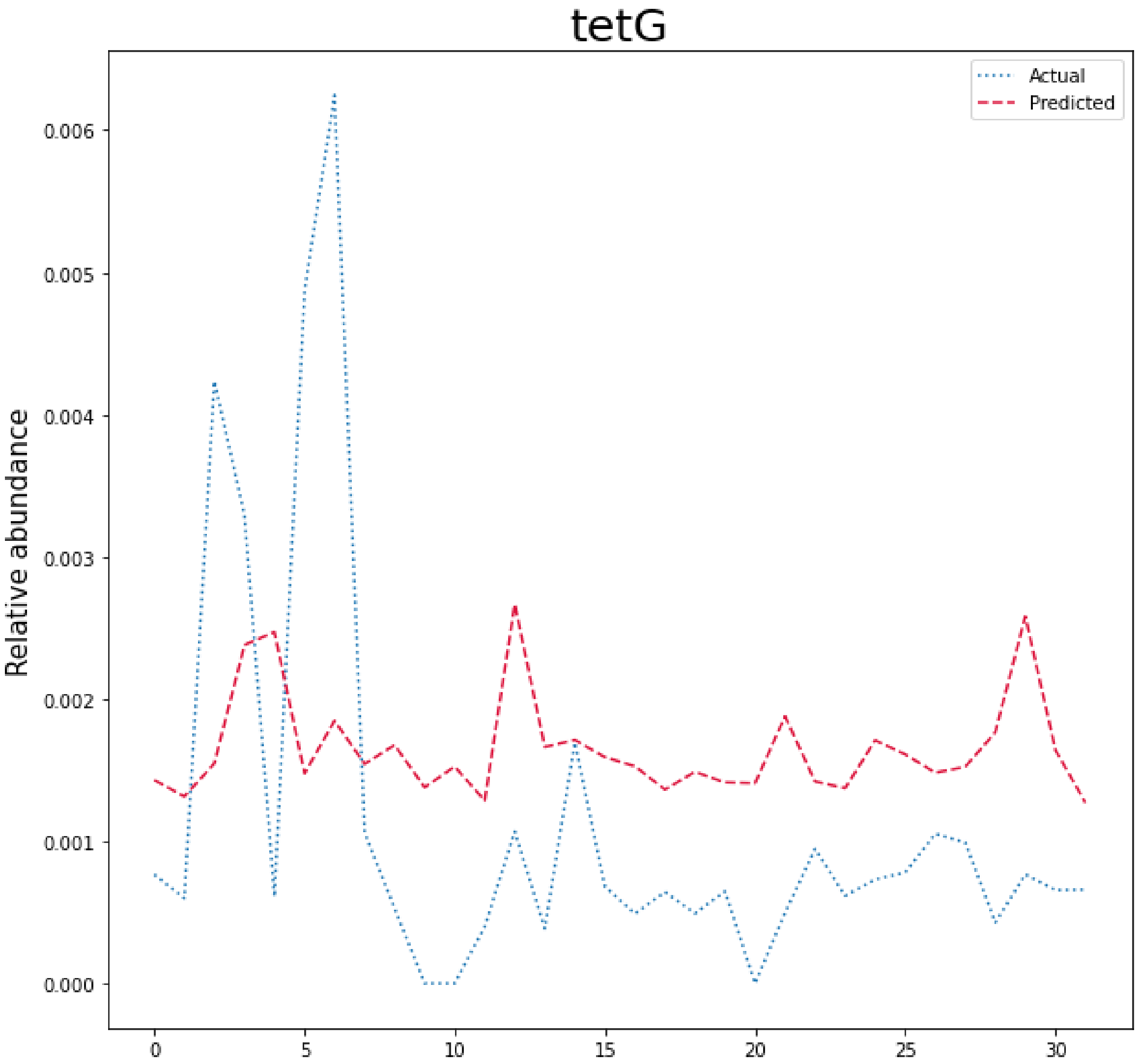
Method (continued)

- Neural network
  - RNN
- Criterion
  - Loss function: MSE
- Evaluation
  - R-square negative



Result: Neural Network





Future Task

- Environmental data as feature
  - Temperature, pH level etc
- Model for other genes
  - genes with low frequency
  - Genes present in Effluents only

Questions