Toxicogenomics-based Water Risk Assessment by Consensus Clustering on Contaminated River

Northeastern Undergraduate Matthew Greenlaw
Mentors: PhD Candidates Sheikh Mokhlesur Rahman and Xin Wen
Principal Investigator: April Gu

ABSTRACT

Objective:

- ❑ Consensus Clustering (CC) is an unsupervised learning algorithm that applies self-organizing maps to perform High-Throughput Screening (HTS)
- ❑ By CC, analyze how stress-related proteins of yeast respond to Dan River samples
- ❑ Find similarity groups between the toxicogenomic responses and the locations or times that assays were recorded

Motivation:

- ❑ The Tox21 of the EPA and EU REACH campaigns are initiatives that demonstrate a growing need for enhanced water risk assessment
- ❑ Contemporary assessments are commonly performed *in vivo*, which requires both time and resources for affected ecosystems

SCOPE

The goal of this research is to identify similarities in polluted water samples effectively in both cost and time by running consensus clustering on toxicogenomics

BACKGROUND

Dan River

- ❑ In 2014, tons of coal ash was spilled in the Dan River
- ❑ Toxicity assays were selectively chosen about the water and sediment both upstream and downstream from the spill

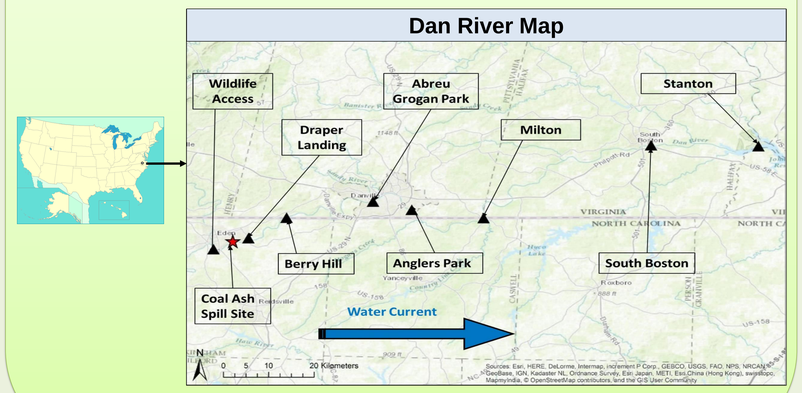


Figure 1: Dan River toxicity assays in Danville, VA.

In Vitro

- ❑ Observes how toxicants could affect microorganisms
- ❑ Yeast commonly appears in environmental water samples; serving as a surrogate to *in vivo* methods for toxicity testing

In Vivo	*In Vitro*
Reliable by directly observing toxicant's effect on target species	Reduces risk uncertainty based on cellular diagnosis
Time-consuming	Time-efficient
Resource-consuming	Resource-efficient

Table 1: Cost-Benefit Analysis of* In Vitro *to* In Vivo *study.

DATA COLLECTION

- ❑ Gene expression was recorded by the rate of GFP fluorescence per cell

$$Gene\ Expression = P = \frac{GFP}{OD}$$

- ❑ The induction factor, *I*, quantifies gene expression alteration per toxicant

$$I = \frac{P_{Experimental}}{P_{Control}}$$

- ❑ Taking the averages of the *I* values among genes gives the net Protein Effect Level Index (PELI), which is then used for consensus clustering

$$PELI = \frac{\int_0^t [|I| - 1]dt}{Exposure\ Time}$$

CONSENSUS CLUSTERING ANALYSIS

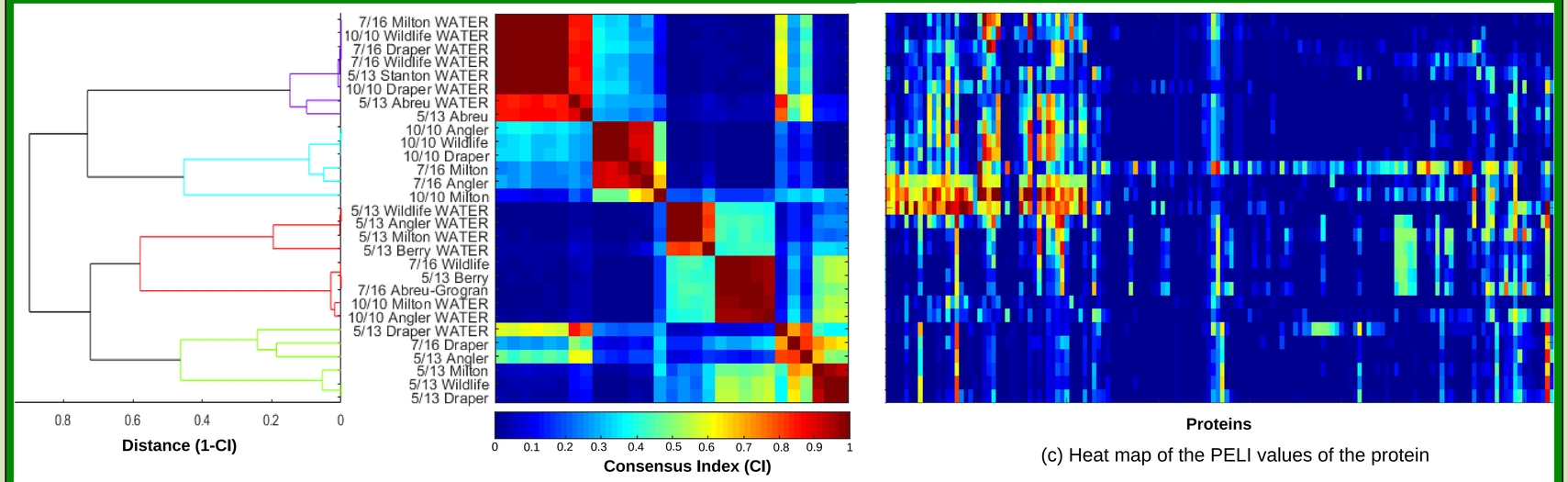
- ❑ Utilizes bootstrapping to increase stability and decrease noise in data
- ❑ Models big data via heat map coloring and hierarchical linkages
- ❑ Uses Self-Organized Maps (SOMs) to determine how well data cluster together with a consensus index (CI)

***i* and *j*: Different Treatments**
***h*: Resampling Index**

Entries	M(*h*) [Connectivity Matrix]	I(*h*) [Indicator Matrix]
0	*i* and *j* are **not** connected	*i* and *j* ***do not*** appear in h^{th} resampling
1	*i* and *j* are connected	*i* and *j* ***do not*** appear in h^{th} resampling

$$CI = \frac{\sum_h M_{(h)}(i,j)}{\sum_h I_{(h)}(i,j)}$$

RESULTS



(a) Dendrogram of the consensus clustering (b) Heat map of the Consensus Index (c) Heat map of the PELI values of the protein

Figure 2: Consensus clustering based on toxicogenomic assay response of the water and sediment samples collected from Dan River. Consensus clustering identifies the similarities of the samples in terms of their toxicogenomic responses.

CONCLUSIONS

- ✓ Dissimilarity exists between protein expression and assay location (i.e. water versus sedimentary samples)
- ✓ Dissimilarity exists between protein expression and assay date of recording
- ✓ More information is needed on samples' chemical composition to infer toxicity levels from microbial protein expression

FUTURE WORK

- ❑ Employ consensus clustering to observe how toxicity potential varies throughout the Dan River
- ❑ Perform *in vitro* assessment for other microbes common in water

Contact Information:
Matthew Greenlaw
<greenlaw.ma@husky.neu.edu>