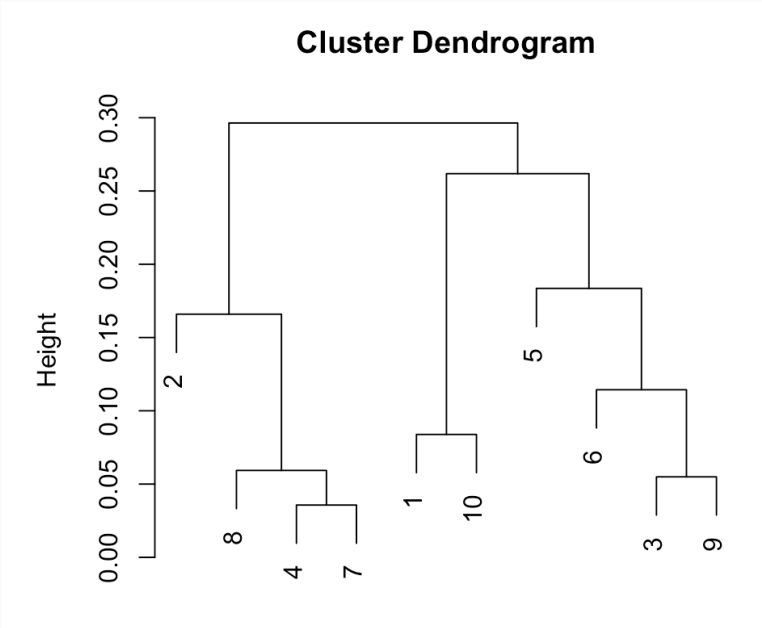
Lag Penalized Weighted Correlation (LPWC)

Thevaa Chandereng, Anthony Gitter

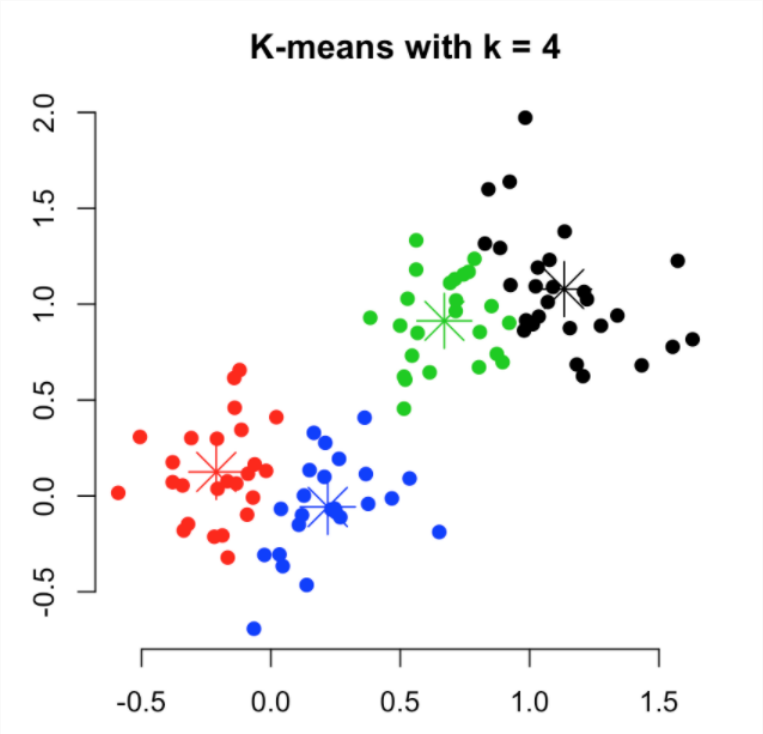
Biological Time Series

- Snapshot of biological functions over time
- Study complex and dynamic biological systems
- Tracking levels of genes/proteins reveals interactions
- Biological time series are shorter compared to time series data in other domains (5-30 time points)
- Similarity in temporal behavior may correspond to similarity in biological processes

Types of Clustering Algorithms



(a) Hierarchical-based clustering



(b) Partition-based clustering. Image adopted from R package vignette factoextra.

Figure 1: Two different clustering strategies

Toy Example: Intuitive Clustering

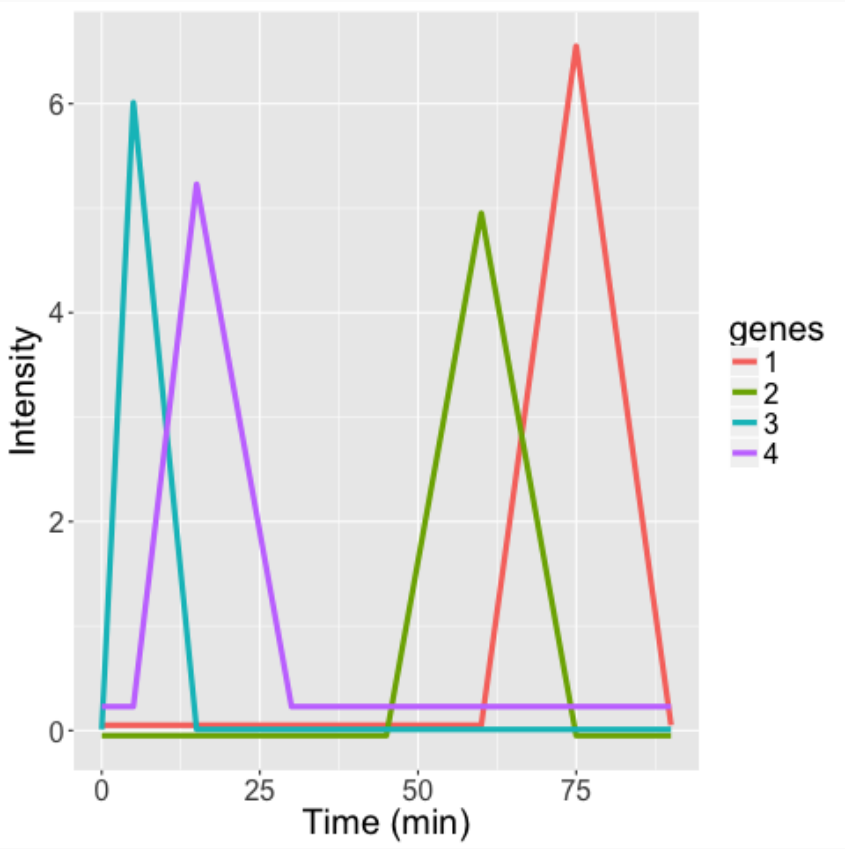
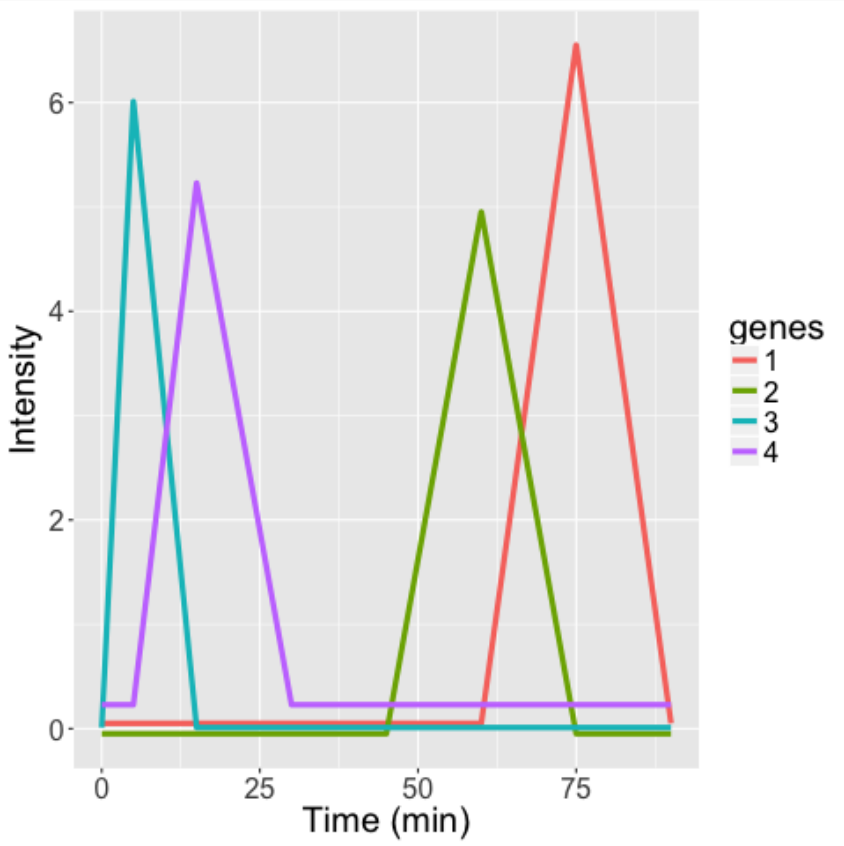


Figure 2: Hypothetical example with 4 genes

Toy Example: Algorithmic Clustering



(a) Hypothetical example with 4 genes

Clustering Algorithm	Cluster 1	Cluster 2
hLPWC/lLPWC	●●	●●
DTW	●	●●●
STS	●	●●●
heuc	●	●●●

(b) Cluster assignment of the 4 genes

Figure 3: Existing methods do not group early and late genes

Motivation

Irregular time sampling

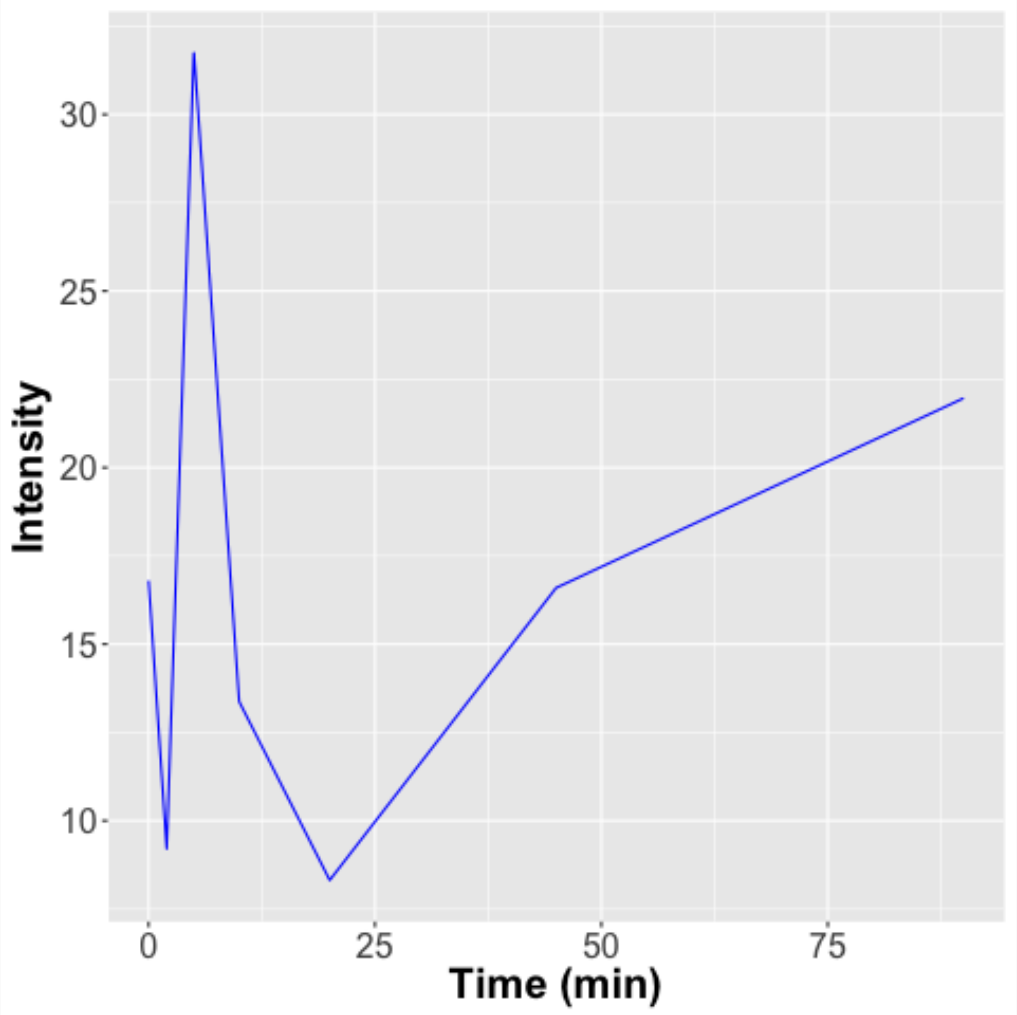


Figure 4: Irregularly sampled time series data

Motivation

Delayed response (lags)

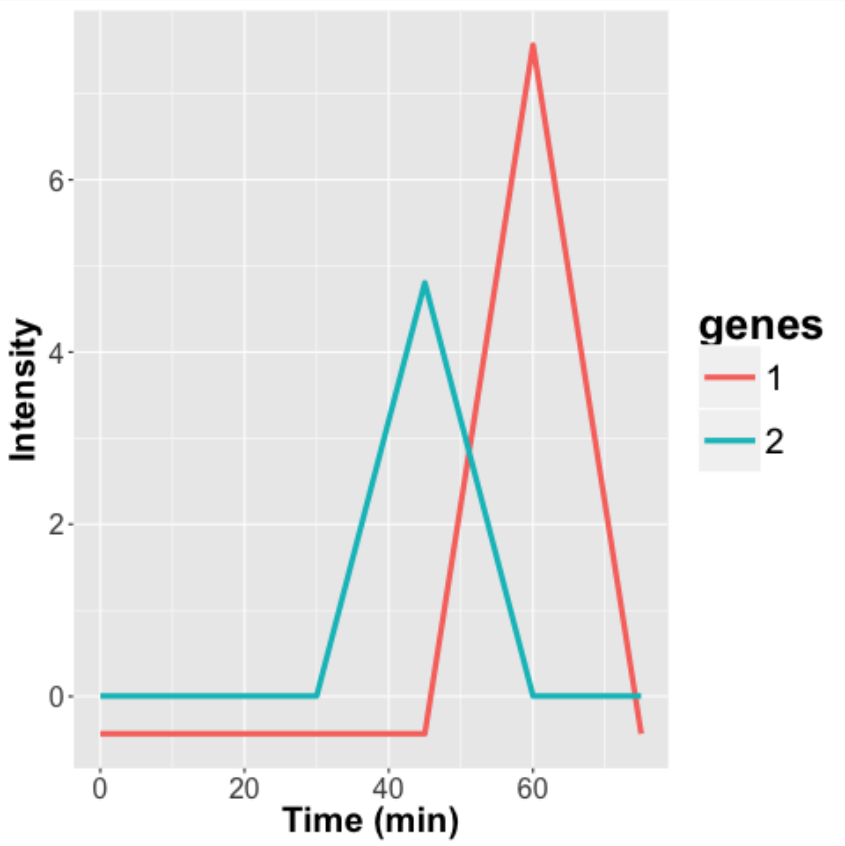


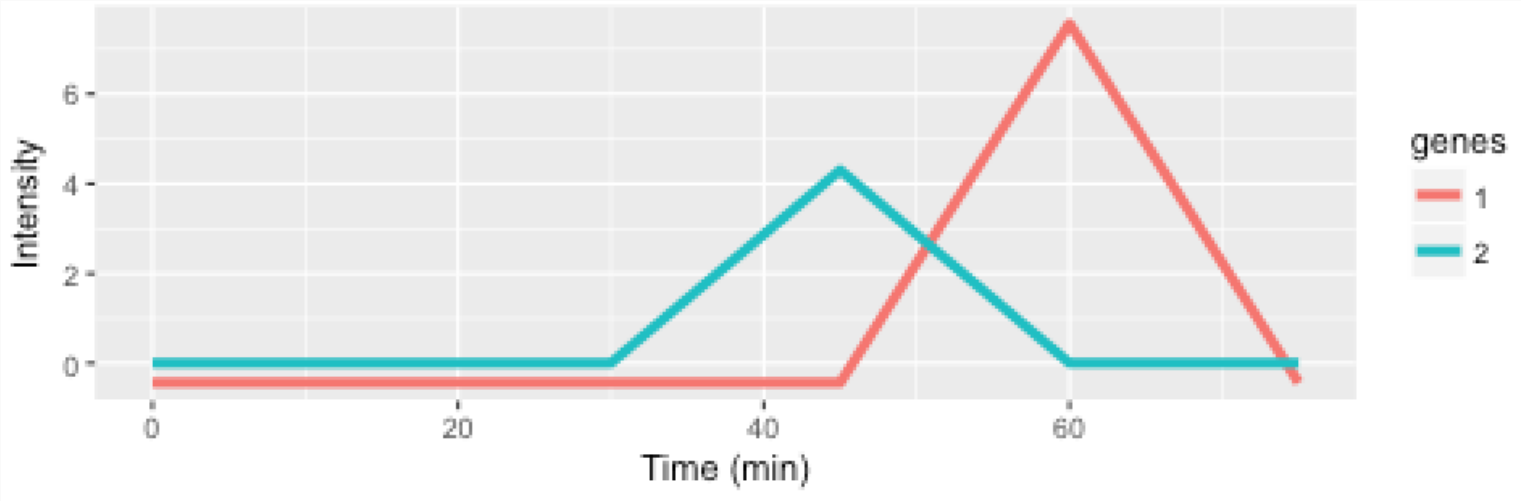
Figure 5: Gene 1 spikes after gene 2

General Formula

The general formula of LPWC

$$corr_{LPWC} = \text{penalty} * \text{weighted correlation}$$

What is a Lag? No Lag Case



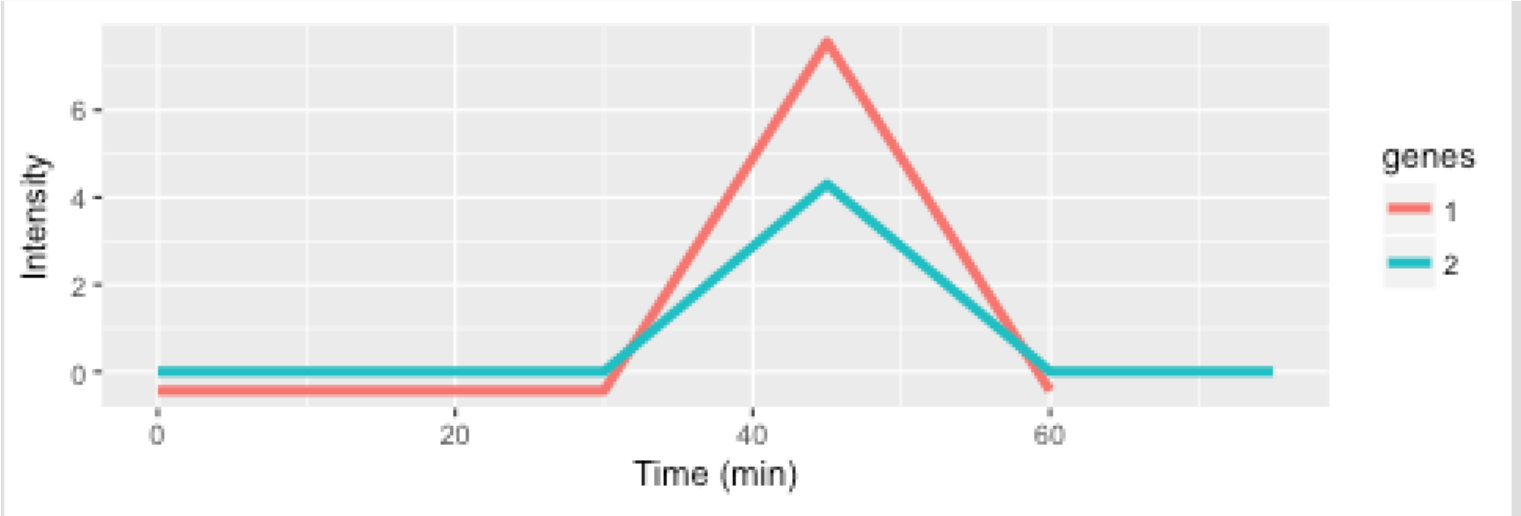
(a) Intensity alignement for no lags case.

Gene 1	0	5	15	30	45	60	75
Gene 2	0	5	15	30	45	60	75

(b) Temporal alignment with the matching intensity in Figure 6a

Figure 6: Gene 1 and gene 2 are not lagged.

What is a Lag? One Lag Case



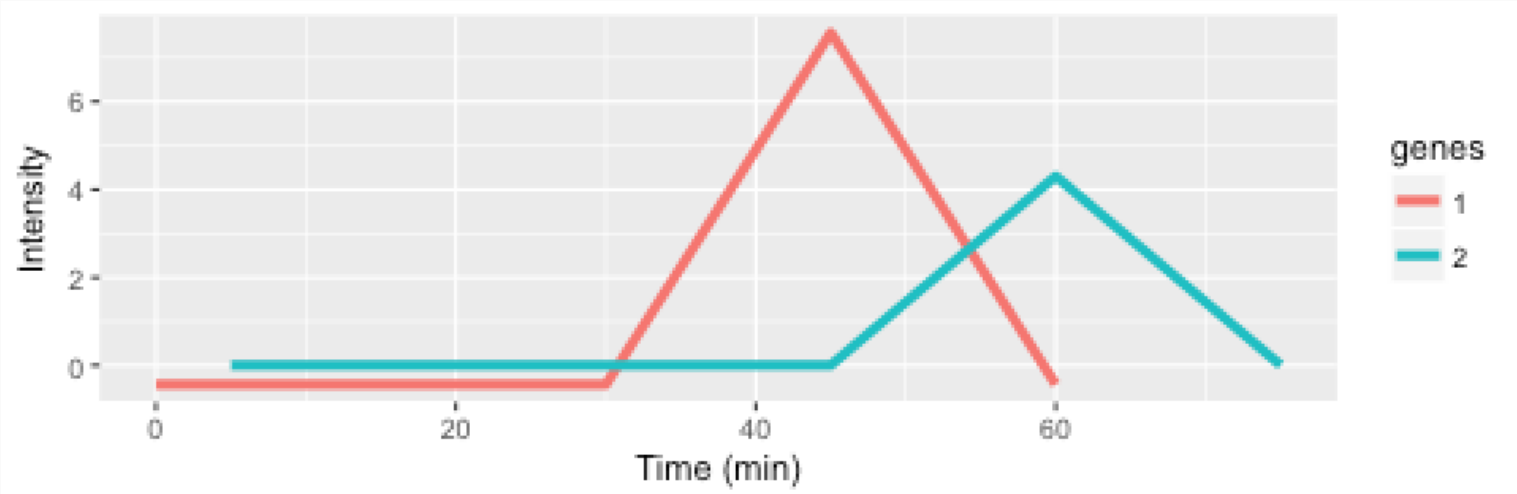
(a) Intensity alignement for gene 1 with lag -1 and gene 2 with no lags.

Gene 1	5	15	30	45	60	75
Gene 2	0	5	15	30	45	60

(b) Temporal alignment with the matching intensity in Figure 7a

Figure 7: Gene 1 with lag -1 and gene 2 with no lags.

What is a Lag? One Lag Case



(a) Intensity alignement for gene 1 with lag -1 and gene 2 with lag 1.

Gene 1	5	15	30	45	60
Gene 2	0	5	15	30	45

(b) Temporal alignment with the matching intensity in Figure 7a

Figure 8: Gene 1 with lag -1 and gene 2 with lag 1.

Method Overview

LPWC is composed of two steps:

- computing optimal lags for each gene
- computing final correlation matrix for all gene

General Formula

$$corr_{LPWC}(i, j, X_i, X_j) = \underbrace{exp(\frac{-E(w)}{C})}_{\text{penalty}} * \underbrace{corr_w(L^{X_i} Y_i, L^{X_j} Y_j, exp(\frac{-w}{C}))}_{\text{weighted correlation}}$$

$$w = (L^{X_i} T_i - L^{X_j} T_j)^2$$

Clustering Accuracy

- Adjusted Rand Index (ARI): similarity between two data clusterings and adjusted for chance
- ARI score close to 1 indicates similar clusterings, score close to 0 otherwise

Simulated Data

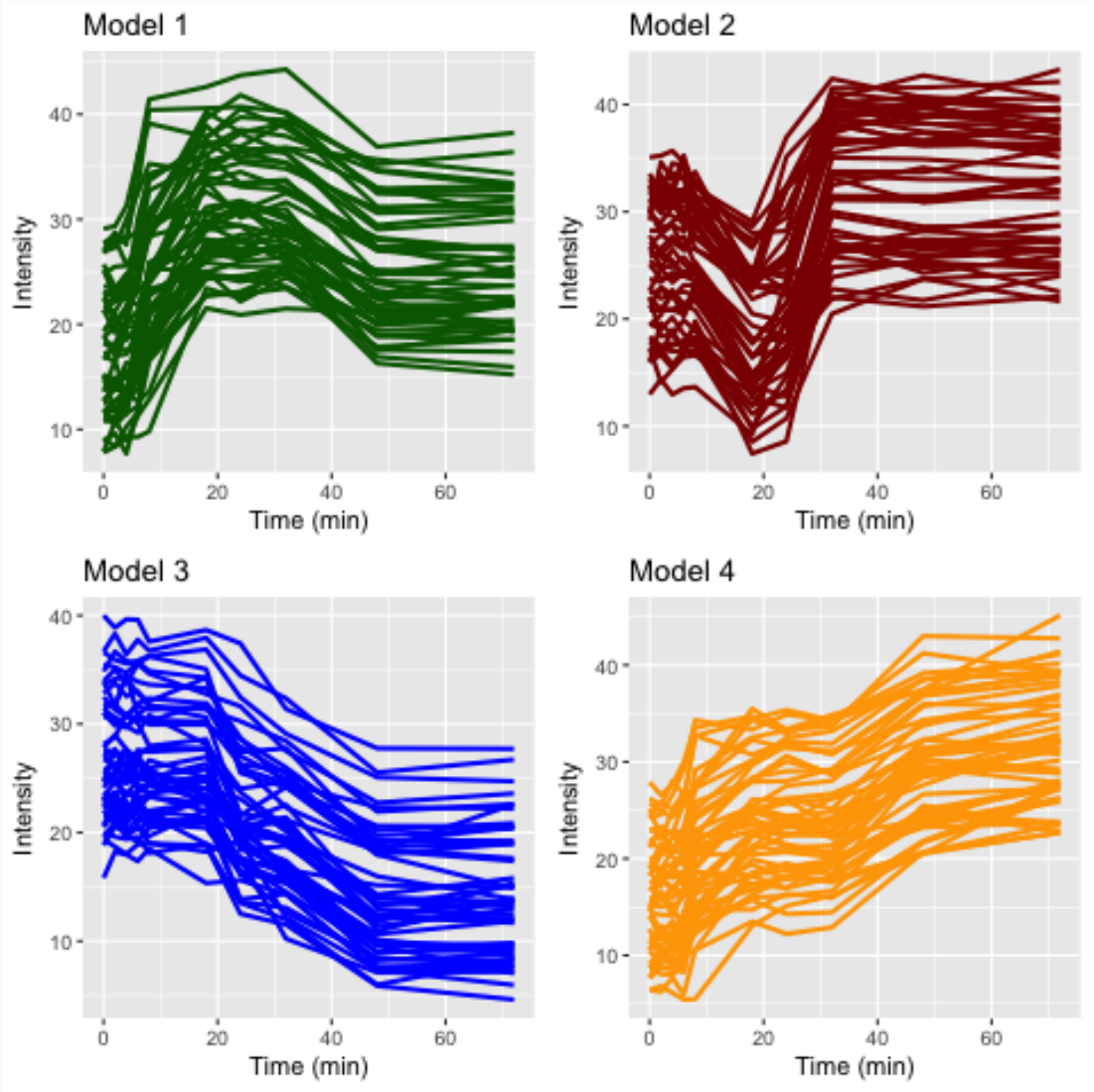


Figure 9: Four models simulated using ImpulseDE. Random noise was added to the model parameters to induce variation around a common trend.

ARI Score for Simulated Data

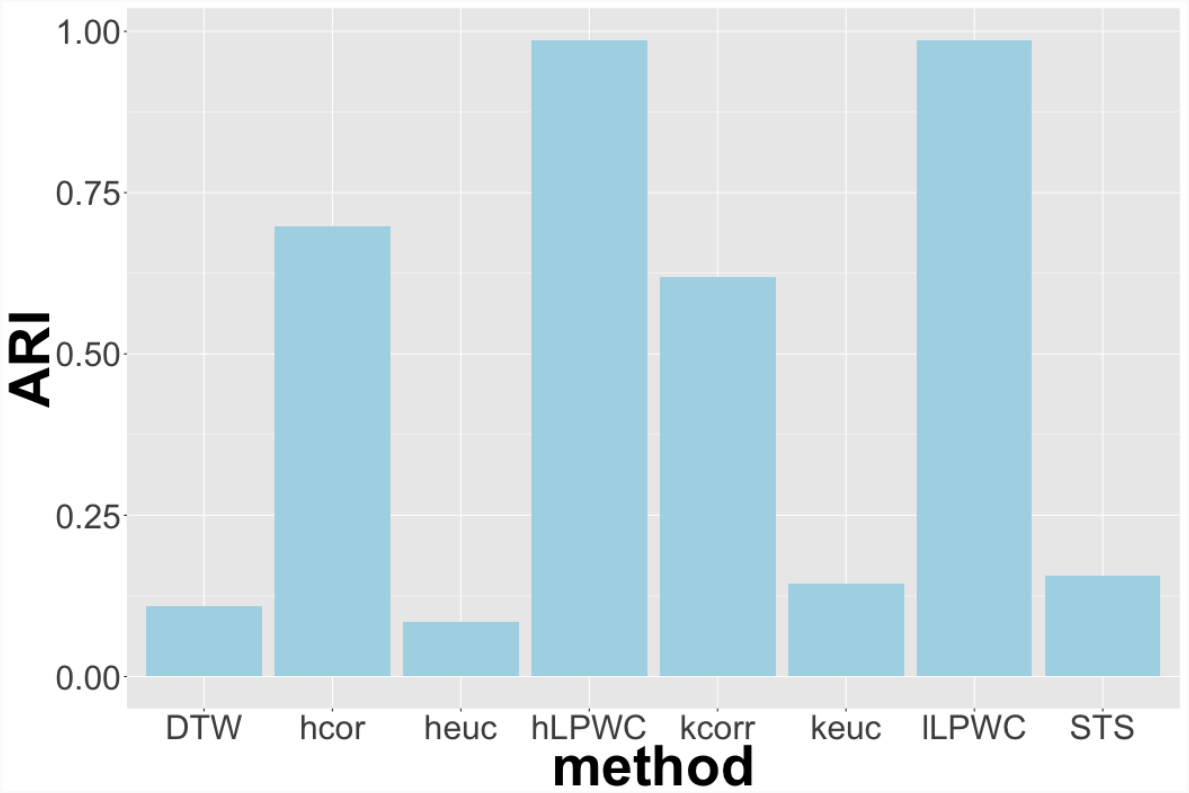


Figure 10: ARI score for different clustering methods for the simulated data where the real clusters are known.

Yeast Osmotic Stress Response Data

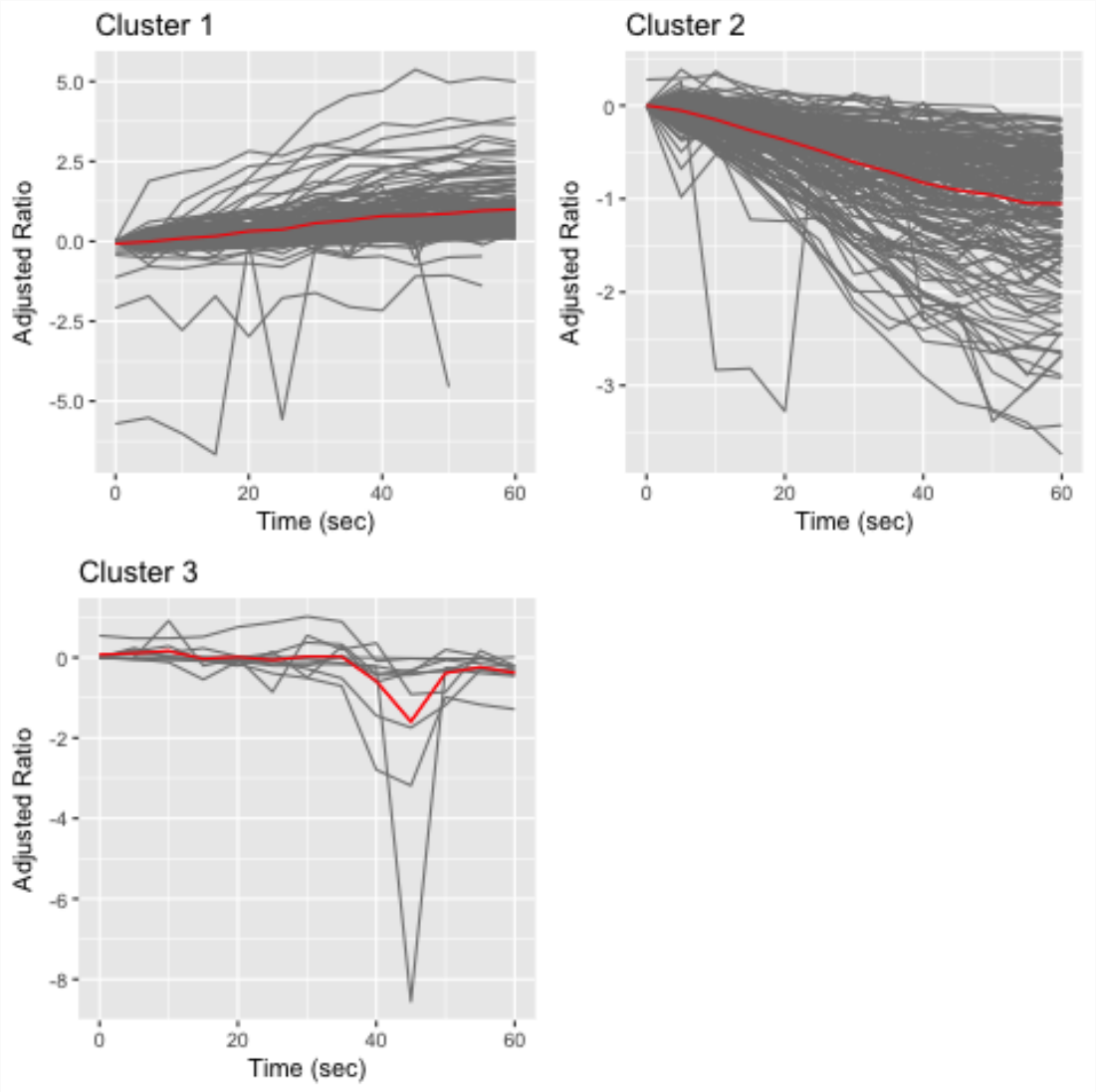


Figure 11: Clustering 344 phosphopeptides in yeast osmotic stress into 3 different clusters.

Conclusion & Future Work

- Algorithm tackles the issue of irregular time samples and delayed responses
- R package available on CRAN (LPWC) and preprint on bioRxiv
- Allow missing data (imputation) and support mixed dataset with different time points
- Improve the optimal lag assignments

Acknowledgements

- Ron Stewart, Karl Broman, James Dowell, Wenzhi Cao, Jen Birstler, and members of Gitter lab
- Funding from the NSF and UW Carbone Cancer Center