Lag Penalized Weighted Correlation (LPWC)

Thevaa Chandereng, Anthony Gitter

Biological Time Series

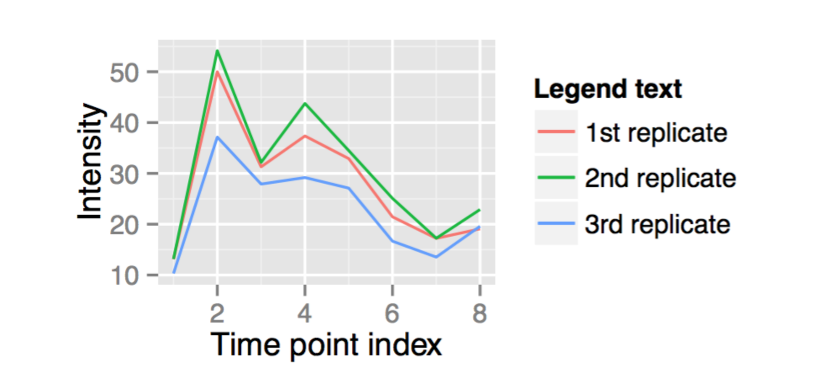


Figure 1: Simple time series plot with 8 time points and 3 replicates

Toy Example: Intuitive Clustering

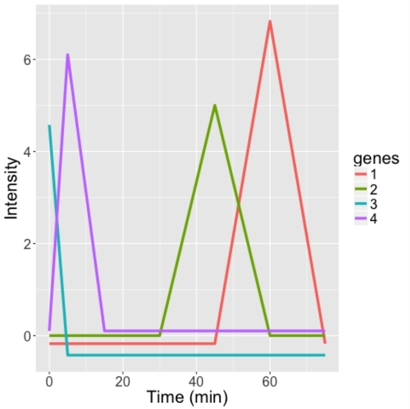
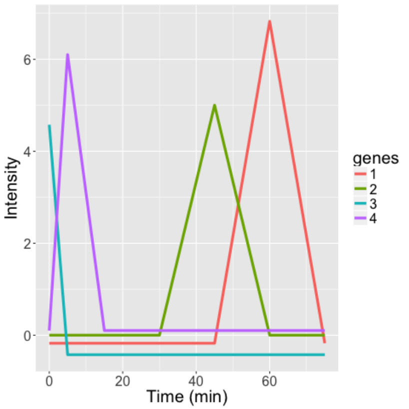
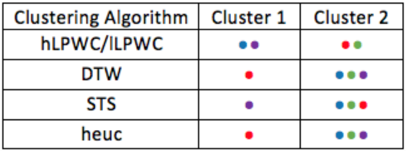


Figure 2: Hypothetical example with 4 genes

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

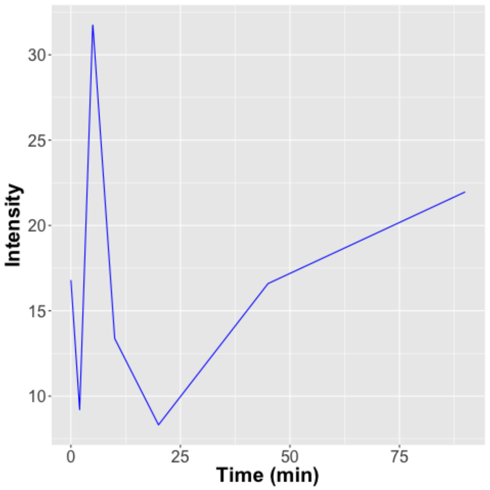


Figure 4: Irregularly sampled time series data

Motivation

Delayed response (lags)

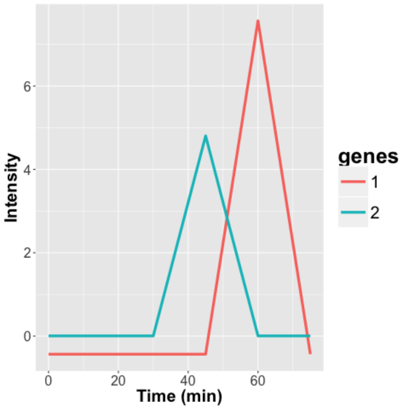


Figure 5: Gene 1 spikes after gene 2

What is a Lag?

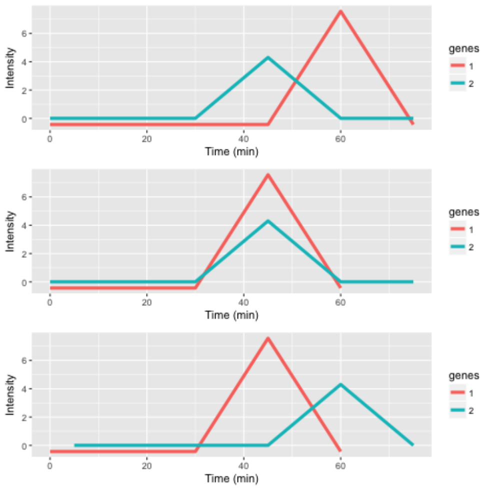


Figure 6: An example of the effects of applying different lags to genes 1 and 2. Gene 1 and 2 are not lagged in top row. Gene 1 with lag 1 i

Method Overview

LPWC is composed of two steps:

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

General Formula

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

Algorithm

Computing optimal lag

$$score_j = \max_{X_i \in \{-m,\ldots,m\}} corr_{LPWC}(i, j, X_i, 0) \quad \forall j \neq i$$

$$lag_j = \arg\max_{X_i \in \{-m,\ldots,m\}} corr_{LPWC}(i, j, X_i, 0) \quad \forall j \neq i$$

Then, a best lag $\hat{X}_i$ for gene i assigned by

$$\hat{X}_i = \arg\max_{k \in \{-m,\ldots,m\}} \sum_{j \neq i} I(lag_j = k) * score_j$$

This is repeated to select a best lag for all genes.

Computing final correlation matrix

$$corr_{LPWC}(i, j, \hat{X}_i, \hat{X}_j) = exp(\frac{-E(w)}{C}) * corr_w(L^{\hat{X}_i} Y_i, L^{\hat{X}_j} Y_j, exp(\frac{-w}{C}))$$

Existing Time Series Clustering Methods

Partition-based

- Short Time-series Expression Miner (STEM)
- Graphical Query Language (GQL)
- Cluster Analysis of Gene Expression Dynamics (CAGED)

Hierarchical-based

- Dynamic Time Warping (DTW)
- Short Time Series Distance (STS)

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

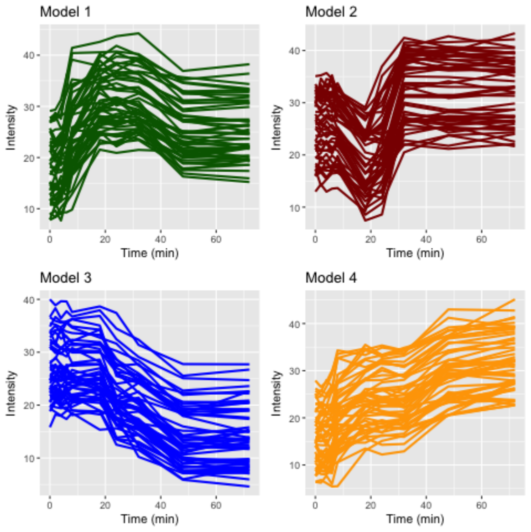


Figure 7: Four models simulated using ImpulseDE. Random noise was added to the model parameters to induce variation around a common trend.

ARI Score for Simulated Data

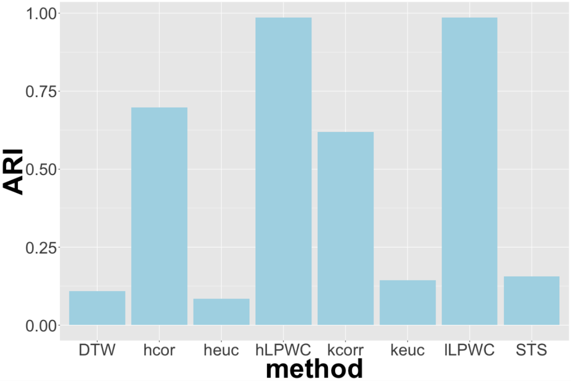


Figure 8: ARI score for different clustering methods for the simulated data where the real clusters are known.

Yeast Osmotic Stress Response Data

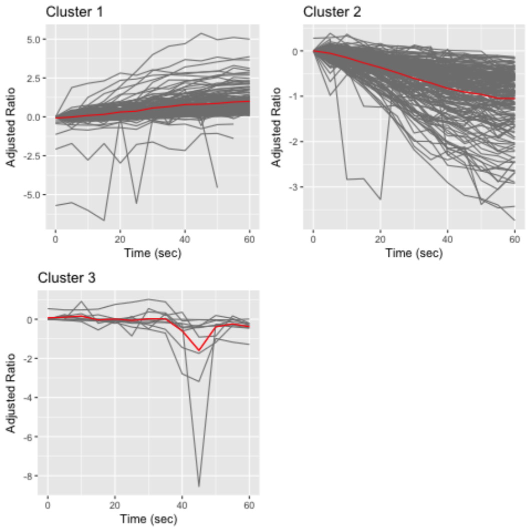


Figure 9: Clustering 344 phosphopeptides in yeast osmotic stress into 3 different clusters.

Conclusion & Future Work

- Algorithm tackles the issue of irregular time samples and delayed responses
- Preference for distance-based or correlation-based clustering is subjective
- 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