Rhythmic clustering of curvy genes

Philip Law

P.J.Law@warwick.ac.uk

Supervisors: Andrew Mead and Vicky Buchanan-Wollaston

THE UNIVERSITY OF WARWICK

SYSTEMS BIOLOGY DOCTORAL TRAINING CENTRE

Introduction

§ **Time series microarray experiments** are often used for observing changes in gene expression levels over time and in response to various treatments

§ Sampling is destructive (i.e. **cross-sectional data**), so there is no autocorrelation between observations at adjacent times

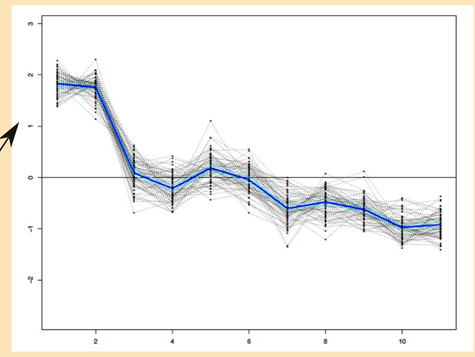
§ Typically clustering is done using programs such as Splinecluster. However, replicate data is not used, genes are not clustered in a biologically meaningful manner, and there is no way to divide the clusters by specific aspects of the curves. In addition, such methods require the expression to be the same across the entire time course

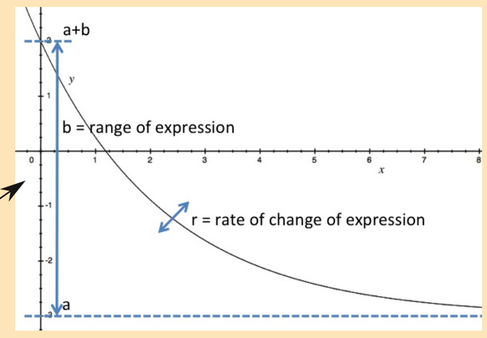
§ Using **linear and nonlinear regression**, observed responses were analysed and genes grouped together based on the shape of the response and the fitted parameters. These parameters can be related to some **physically interpretable process**

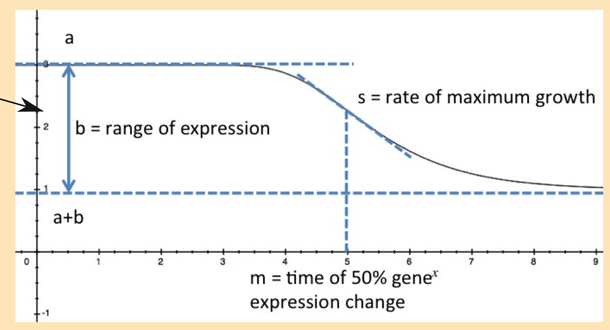
§ Thus it is possible to identify values such as **critical time points** (e.g. start of differential expression and the maximum or minimum expression), or the **maximum rate of change**

§ The overall aim is to develop a statistical analysis approach to model the relationships between genes to predict the impacts of multiple stresses

§ Investigated the PRESTA long day senescence data







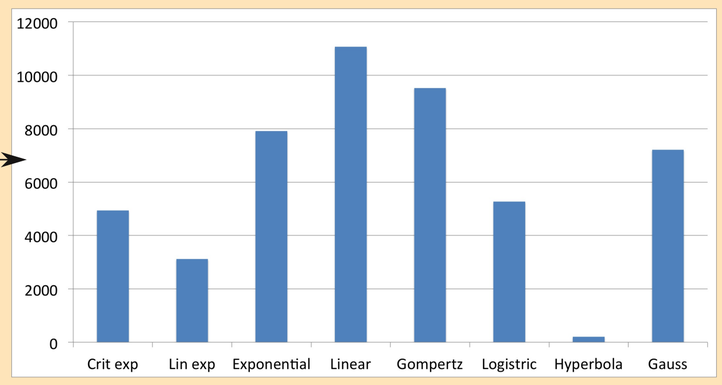
Gene expression → Fit models

-Critical exponential
-Exponential
-Gaussian
-Gompertz
-Hyperbola
-Linear
-Line+exponential
-Logistic

→ Use AIC (Akaike Information Criterion) to find best model(s)

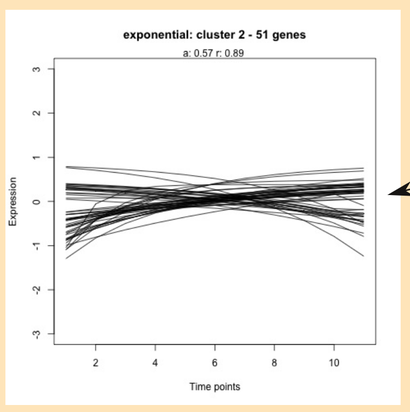
→ Use goodness of fit techniques to determine if good fit
- Adjusted R^2
- Lack-f-fit adjusted R^2
- F test

Distribution of model fits



Clusters parameters with 90% similarity

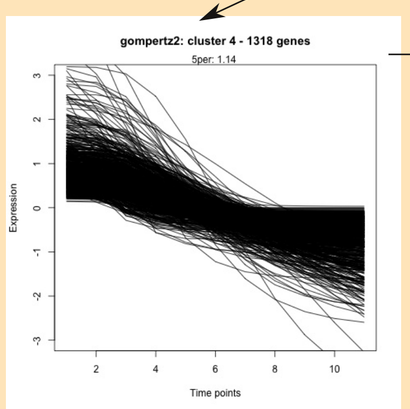
Cluster on *r* (rate of change) parameter of exponential



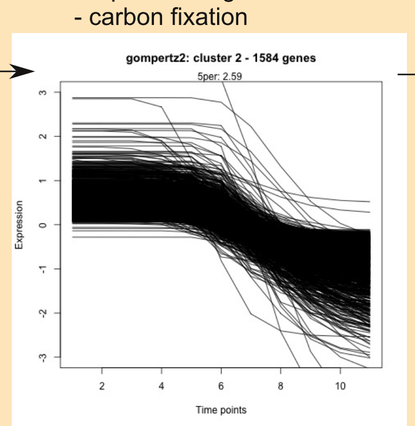
- oxidoreductase activity
- structural constituent of ribosome
- translation
- carbohydrate metabolic process

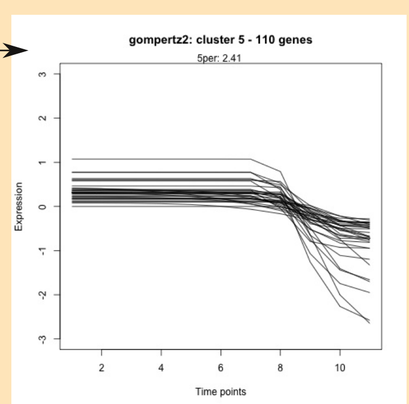
Can also cluster parameters on infered parameters, such as time of 5% to maximum growth of Gompertz. It is thus possible to identify genes that change at certain times. Below shows a series of genes changing at different times

- photosynthesis
- carbohydrate metabolism
- reponse to light
- carbon fixation



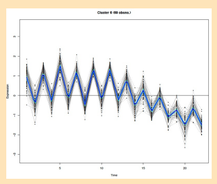
- amino acid activation
- regulation of cell growth
- various biosynthetic and metabolic processes



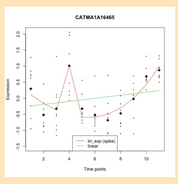


- No over represented GO terms

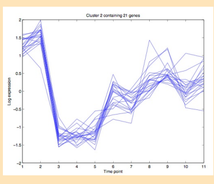
Why did some genes not fit well?
- Oscillating behaviour (left)
- "Spike" response (middle)
- Model shape not included (right)



-Fourier transform
-Sign test



-Leave one out regression
-Find points of high leverage



-Compound model
-New models

Conclusions Future Work

§ Regression modelling allows the consideration of a range of response shapes, the comparison of biologically-interpretable parameter estimates, and the assessment of the goodness-of-fit of different models

§ Grouping genes based on the best-fitting model shape and the similarity of the fitted parameter values could lead to insights into the relationships between genes, such as genes that are activated at a particular time

§ Future work includes developing a method to estimate the "true" biological time associated with a gene expression response
- Identify samples taken at the same time but were at different developmental stages, so belong to an earlier/later timepoint

§ Link parameters to the associated biological processes to possibly provide a better assessment of the timing of plant stress responses

§ Compare control/treated aspects of the same experiment
- Genes with same shapes, but different parameters; or genes with different shapes between treatments

§ Several other sets of time series data from the PRESTA project that will also be analysed, including *Botrytis cinerea* infection; *Pseudomonas syringae* virulent and non-virulent infection; drought; and high light

§ In addition, various multivariate statistical approaches will be applied to link gene expression responses with other plant responses, and thus predict the effect of multiple stresses