Nonlinear regression as a method for gene expression profile analysis

1. Introduction

- Time series microarray experiments are a popular technique for observing the changes of gene expression levels, over time and in response to some treatment
- This project used nonlinear regression analyses to model the observed responses and grouped genes based on the shape of the response and the fitted parameters. A GO analysis was used to determine potential functional characteristics of groups of genes
- Nonlinear regression provides more information about the observed response than current analysis methods, such as physically interpretable model parameters, as well as various statistics to determine how well a set of genes fits a particular shape

2. Algorithmic development

- Test data were obtained from 2 Arabidopsis experiments: senescence (11 timepoints), and Botrytis infection experiments (16 timepoints). Botrytis dataset contained mock and treated samples. Data was cross sectional, where biological replicates are from different organisms (plants)
- Biological replicate values produced from MAANOVA analysis of raw data
- 9 models used – linear, quadratic, cubic, exponential, critical exponential, linear+exponential, logistic, Gompertz, Michalis-Menten
- First 3 models are linear, rest nonlinear – model list is not exhaustive. Models selected as the most commonly occurring as seen in a spline clustering. Possible to add additional models
- Programmed using R, Python, RPy

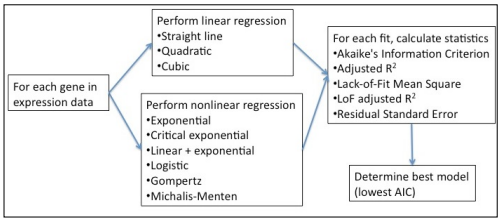

Figure 1: Finding the best model
To find the best model, all the models were fitted to the expression data for each gene. The model with the smallest AIC was selected as the best

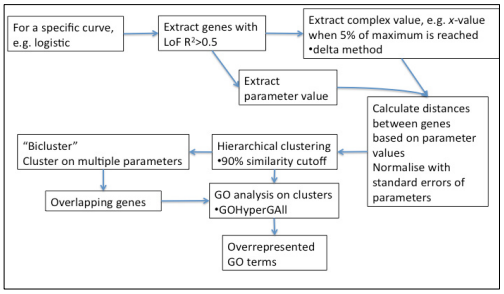

Figure 2: Clustering the models by parameter values
Once the best fits in a specific model had been indentified, the significant fits could be extracted, using a lack-of-fit (LoF) adjusted R^2>0.5, an arbitrarily selected threshold. Comparing the parameters of the model fits of these genes, it was possible to produce clusters, using hierarchical clustering.

Bi-clustering identifies groups of genes that falling into distinct clusters for two different parameters, and can provide more information about the response. These clusters could then be analysed for over-represented GO terms

3. Senescence results

These results focus specifically on the genes that had a logistic shaped expression profile (logistic equation shown on right)

$$y = a + \frac{b-a}{1+exp((xmid-x)/scal)}$$

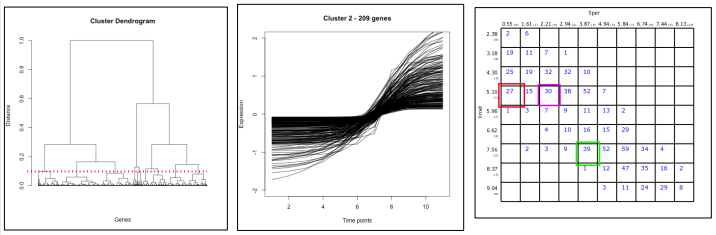

Figure 4: Dendogram, clusters and biclusters
Shown is a dendogram with the 10% dissimilarity cutoff, and a cluster after clustering by the midpoint parameter (*xmid*). The bicluster highlights sets of genes that are influenced by two parameters. This is advantageous as it relates possibly diverse model parameters together.
Clustering by *xmid* identified sets of genes that were activated at different times, with terms early on relating to water deprivation, eventually leading to cell wall degradation.

4. Botrytis results

With the Botrytis data, the shapes between treatments were compared graphically.

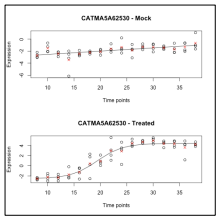

Figure 5: Different models
Sometimes the responses were very different, in this case mock was linear and treated was logistic. This gene encodes a MAP kinase known to be involved in response to pathogens.

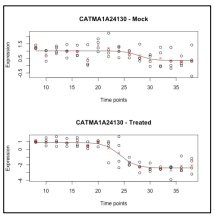

Figure 6: Similar but different
The responses could have the same shape (both logistic), but different parameters. The gene encodes a transcription factor

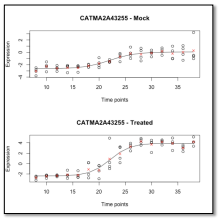

Figure 7: Similar but different 2
The responses could also have the same shape (both logistic) as well as similar model parameters. The two responses here were almost identical save for the maximum asymptote parameter.

5. Concluding discussion

The use of a regression modelling approach allows the consideration of a range of response shapes, comparison of biologically-interpretable parameter estimates, and assessment of the goodness-of-fit of different models. Grouping of genes based on the best-fitting model shape and the similarity of one or more fitted parameter values (or derived parameters) can lead to insights into the relationships between genes. In addition, investigating individual parameter values can help to identify particular characteristics of groups of genes, such as those activated at a particular time

SYSTEMS BIOLOGY DOCTORAL TRAINING CENTRE

Philip Law
P.J.Law@warwick.ac.uk

Supervised by:
Andrew Mead
Katherine Denby
Vicky Buchanan-Wollaston

THE UNIVERSITY OF WARWICK