Bayesian Hierarchical Clustering for Microarray Time Series Data with Replicates and Outlier Measurements

Emma J. Cooke, Paul D. W. Kirk, Richard S. Savage, Robert Dawkins and David L. Wild

email: e.cooke@warwick.ac.uk, p.kirk@warwick.ac.uk, r.s.savage@warwick.ac.uk and d.l.wild@warwick.ac.uk

To be released as part of the BHC Bioconductor package: http://www.bioconductor.org/packages/release/bioc/html/BHC.html

1. Introduction

✸Grouping together genes which have similar expression values measured by a microarray can suggest genes which are coregulated during a particular biological process.

✸We present a model-based Bayesian hierarchical clustering algorithm for microarray time series data that employs Gaussian process regression to capture the structure of the data. Our method can model outlier and replicate values, learns the optimum number of clusters given the data and can incorporate non-uniformly sampled time points.

✸Our time series clustering algorithm has been submitted to the R Bionconductor suite as an extension to the BHC package and will be available in future Bioconductor releases.

2. Method

✸We model the gene profiles as being drawn from a Gaussian distribution with mean function *f* and covariance function *K*.

✸We have implemented both squared exponential (BHC-SE) and cubic spline (BHC-C) covariance functions.

✸The log marginal likelihood of the data is:

$$\log\, P(\boldsymbol{y_c}|H_1^c) = -\frac{1}{2}(\boldsymbol{y_c}^T K^{-1}\boldsymbol{y_c}) - \frac{1}{2}(\log|K|) - \frac{N}{2}(\log(2\pi))$$

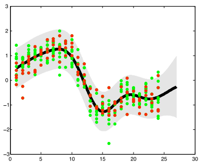
✸ ***y$_c$*** = all the data in cluster *C*. *H$_1$^{c}* = hypothesis that all the data in cluster *C* were generated from the same Gaussian process.

✸Initially, each gene begins in its own cluster. At each stage the two most similar clusters are merged together. The BHC algorithm (Heller and Ghahramani 2005), uses Bayes rule to find *r$_c$*, the posterior probability that two clusters are merged:

$$r_c = \frac{\pi_c P(\boldsymbol{y_c}|H_1^c)}{P(\boldsymbol{y_c}|\text{all tree-consistent ways of partitioning data})}$$

✸Observation = signal + noise

$$y(t_i) = f(t_i) + \varepsilon,\ \varepsilon \sim N(0, \sigma_\varepsilon^2)$$



✸Gene expression *vs* time, merged red and green clusters.

✸The pair of clusters with the highest *r$_c$* are merged at each stage. When *r$_c$* < 0.5 for all cluster pairs, BHC has found the optimum number of clusters.

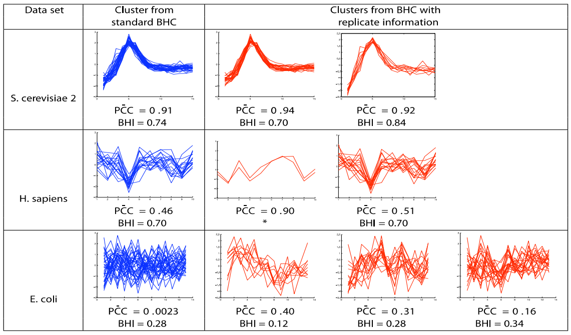
✸We have also implemented the option for BHC to use a mixture model likelihood which allows a small proportion of the data to be modelled as outliers.

✸ We assume the single data point *y$_n$* has been generated from an outlier likelihood function *B$_n$* with a small probability *b* and generated from a Gaussian process likelihood function *A$_n$* with probability *a* = 1-*b*. Then the likelihood function for a cluster with *N* data points is:

$$P(\boldsymbol{y}|\boldsymbol{f},\boldsymbol{\theta}) = \prod_{n=1}^{N}(aA_n + bB_n)$$

4. Including replicate information

✸Time series BHC optionally allows replicate information to be used in order to inform a prior distribution of the noise variance. Using replicate information can split a noisy cluster into smaller, less noisy clusters.



Data set	Cluster from standard BHC	Clusters from BHC with replicate information		
S. cerevisiae 2	PC̄C = 0 .91 BHI = 0.74	PC̄C = 0 .94 BHI = 0.70	PC̄C = 0 .92 BHI = 0.84	
H. sapiens	PC̄C = 0 .46 BHI = 0.70	PC̄C = 0 .90 *	PC̄C = 0 .51 BHI = 0.70	
E. coli	PC̄C = 0 .0023 BHI = 0.28	PC̄C = 0 .40 BHI = 0.12	PC̄C = 0 .31 BHI = 0.28	PC̄C = 0 .16 BHI = 0.34

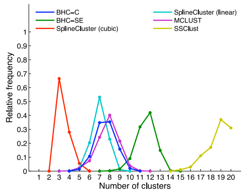
3. Comparison to other methods

✸We used 4 time series microarray data sets for analysis. The table below shows the number of genes, time points, replicates and run times on a single core 2.40 GHz Intel Xeon CPU, for these data sets.

Data set	BHC-SE	BHC-SE mixture model	Genes	Time points	Replicates	Reference
S. cerevisiae 1	6m 3s	38m 49s	169	17	N/A	Cho *et al.* 1998
S. cerevisiae 2	24m 8s	5h 48m	440	15	2	Orlando *et al.* 2008
H. sapiens	19s	49s	58	10	44	Rangel *et al.* 2004
E. coli	7m 6s	34m 39s	200	13	6	Carzaniga *et al.* 2007

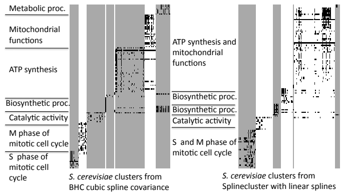
✸To measure the quality of a clustering we used the Biological Homogeneity Index (BHI) (Datta and Datta 2005), which uses gene ontology (GO) annotations to assess the biological homogeneity of the clusters. We also used the average Pearson Correlation Coefficient (PCC) to measure the similarity of gene profiles within clusters. Higher scores are better in both cases.

✸We clustered the 4 datasets above with BHC-SE and BHC-C and compared the clustering with other methods: SplineCluster with linear and cubic splines (Heard *et al.* 2005), SSClust (Ma *et al.* 2006) and MCLUST (Yeung *et al.* 2001). BHC gives the highest BHI and PCC scores in all cases.



✸We generated 1000 data sets from 13 synthetic clusters. BHC-SE does better than the other methods at recovering the correct number of clusters.

✸Over-represented GO annotations for the BHC-C clusters (BHI = 0.73) and the SplineCluster clusters using linear splines (BHI = 0.69), using the *S. cerevisiae 1* data set.



S. cerevisiae clusters from BHC cubic spline covariance

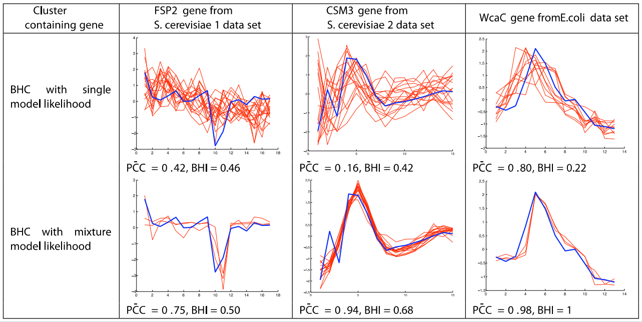
S. cerevisiae clusters from Splinecluster with linear splines

✸BHC-C is able to separate the clusters of mitochondrial and ATP synthesis functions and also the M and S Phase mitotic cell cycle genes, that SplineCluster combines together.

5. Modelling outliers

✸Using a mixture model likelihood can allow genes with noisy observations to be grouped with other genes of similar biological function. The mixture model treats:

✸FSP2 gene time point 11 as an outlier
✸CSM3 gene time point 2 as an outlier
✸WcaC gene time point 4 as an outlier



Cluster containing gene	FSP2 gene from S. cerevisiae 1 data set	CSM3 gene from S. cerevisiae 2 data set	WcaC gene fromE.coli data set
BHC with single model likelihood	PC̄C = 0 .42, BHI = 0.46	PC̄C = 0 .16, BHI = 0.42	PC̄C = 0 .80, BHI = 0.22
BHC with mixture model likelihood	PC̄C = 0 .75, BHI = 0.50	PC̄C = 0 .94, BHI = 0.68	PC̄C = 0 .98, BHI = 1

THE UNIVERSITY OF WARWICK

EPSRC
Engineering and Physical Sciences Research Council