Author(s): Steven J. Kiddle , Oliver P. F. Windram , Stuart McHattie Andrew Mead , Jim Beynon, Vicky Buchanan-Wollaston , Katherine J. Denby and Sach Mukherjee
Article Title: Temporal clustering by affinity propagation reveals transcriptional modules in Arabidopsis thaliana

Year of publication: 2010
http://dx.doi.org/ 10.1093/bioinformatics/btp673
Publisher statement: None

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–8

Temporal clustering by affinity propagation reveals transcriptional modules in *Arabidopsis thaliana*

Steven J. Kiddle [1,4,*], Oliver P. F. Windram [4], Stuart McHattie [1,4], Andrew Mead [4], Jim Beynon [1,4], Vicky Buchanan-Wollaston [1,4], Katherine J. Denby [1,4] and Sach Mukherjee [2,3*]

[1]Warwick Systems Biology Centre, Warwick University, CV4 7AL, UK
[2]Department of Statistics & [3]Centre for Complexity Science, Warwick University, CV4 7AL, UK
[4]Warwick HRI, Warwick University, Wellesbourne, CV35 9EF, UK

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Identifying regulatory modules is an important task in the exploratory analysis of gene expression time series data. Clustering algorithms are often used for this purpose. However, gene regulatory events may induce complex temporal features in a gene expression profile, including time delays, inversions and transient correlations, which are not well accounted for by current clustering methods. As the cost of microarray experiments continues to fall, the temporal resolution of time course studies is increasing. This has led to a need to take account of detailed temporal features of this kind. Thus, while standard clustering methods are both widely used and much studied, their shared shortcomings with respect to such temporal features motivates the work presented here.

Results: Here, we introduce a temporal clustering approach for high-dimensional gene expression data which takes account of time delays, inversions and transient correlations. We do so by exploiting a recently introduced, message-passing-based algorithm called Affinity Propagation (AP). We take account of temporal features of interest following an approximate but efficient dynamic programming approach due to Qian *et al*. (2001). The resulting approach is demonstrably effective in its ability to discern non-obvious temporal features, yet efficient and robust enough for routine use as an exploratory tool. We show results on validated transcription factor-target pairs in yeast and on gene expression data from a study of *Arabidopsis thaliana* under pathogen infection. The latter reveals a number of biologically striking findings.

Availability: Matlab code for our method is available at http://www.wsbc.warwick.ac.uk/stevenkiddle/tcap.html.

Contact: {s.j.kiddle,s.n.mukherjee}@warwick.ac.uk

1 INTRODUCTION

Gene expression analysis by microarrays is now a well established approach in high-throughput biology. Time course studies are widely used to probe the dynamics of gene expression and uncover underlying regulatory programs. As costs per array have continued to fall, the temporal resolution of such studies (in the sense of the number of discrete time points sampled) has increased. Indeed, it is now common to see studies with 20 or more time points over timescales of hours to days. A central task in the exploratory analysis of these high-dimensional time series is that of identifying subsets of genes which are functionally related, for example transcription factors (TFs) and their targets, genes which share a regulatory program and so on. Following much of the recent literature we call such subsets modules (Bar-Joseph *et al*., 2003; Segal *et al*., 2003). Module identification plays a key role both in the generation of experimental hypotheses and in informing subsequent modelling. Microarray data which highlight a set of genes as possibly functionally related can suggest specific follow-up experiments, for example using interventions targeted at module members. Equally, module identification informs further computational work. The inference of gene regulatory networks (e.g. using Bayesian networks or Gaussian graphical models), for example, rapidly grows more challenging in higher dimensions. In the same way, mechanistic models of gene expression (ODE, PDE or statistical mechanical), become much more tractable for small sets of genes. Thus, identifying transcriptional modules can greatly aid downstream, detailed quantitative analysis.

Clustering algorithms are widely used for the purpose of identifying gene modules (e.g. Ghosh & Chinnaiyan, 2001; Heard *et al*., 2005; Thalamuthu *et al*., 2006). Such algorithms seek to partition the set of genes into subsets whose within-subset similarity is high relative to between-subset similarity. The most widely used notions of similarity are simple vector distances between temporal profiles, and include the Euclidean distance, Pearson's correlation coefficient (PCC) and Mahalanobis distance (used in Gaussian mixture models). Loosely speaking, these methods seek to find subsets of genes which *look* similar in the sense of having highly correlated expression profiles. This in turn means that these methods are well suited to detecting modules whose members are co-regulated (Yona *et al*., 2006), for example by a shared TF, and where regulatory events are simultaneous, at least up to the temporal resolution of the dataset.

However, the general strategy of clustering by straightforward profile similarity suffers from a number of drawbacks. First, while

*to whom correspondence should be addressed

it is arguably well suited to certain cases of simultaneous co-regulation, it is not as well suited to finding genes which regulate each other. In these settings there can be a time lag between a change in the profile of the regulator and the corresponding change in its target. At very low temporal resolutions, this may not be an issue, because the changes, if detected, may appear as *de facto* simultaneous. However, at higher temporal resolutions time lags become an important issue; we show experimental examples below.

Second, even when a set of putatively co-regulated genes can be identified, the task of identifying a shared TF remains a challenging one. A widespread approach is to use sequence analysis to discover upstream motifs, shared among module members, which may correspond to TF binding sites. However, even when upstream motifs can be found, TFs that bind to these sequences are often unknown, particularly in higher organisms. This motivates a need for module finding methods which can identify subsets including both regulator and targets directly from expression data.

Third, many existing approaches do not account for transient correlations, in which gene profiles are similar only within a certain time window, and not well correlated outside it. This can arise for example in longer time courses, where the underlying biological process driving profile similarity is itself transient, such that at its end, the genes revert almost to a background level of variation. Two-way clustering or biclustering (Hartigan, 1972; Lazzeroni & Owen, 2002; Balasubramaniyan *et al*., 2004; Madeira & Oliveira, 2005; Meng *et al*., 2009) has been used to address the issue of transient correlations. Here, clusters are sought which form subsets of both genes and (contiguous) time points. However, robust biclustering remains computationally challenging on account of the vast number of possible biclusters that can be formed. Finally, inversions in the sense of negative correlation/co-expression can be important when regulatory relationships are repressive, but are not always accounted for by clustering methods.

In order to account for these temporal features, a natural idea is to carry out cluster analysis using richer similarity measures in place of a simple vector distance; this idea appears several times in the literature (Qian *et al*., 2001; Schmitt *et al*., 2004; Balasubramaniyan *et al*., 2004; Smith *et al*., 2009). However, doing so brings with it a non-trivial computational burden, especially under conditions of high dimensionality and high temporal resolution (and resulting longer time lags). Under Euclidean distance and its variants clusters can be characterized by cluster-level statistics such as the mean; this in turn permits (relatively) fast iterative computation *via* algorithms such as K-means and Expectation-Maximization (EM). In contrast, temporally rich gene-gene similarity measures typically do not give an analogue to cluster mean. The standard approach then is to use an iterative algorithm known as K-centres (or K-medioids) (see e.g. Hastie *et al*., 2001). However, K-centres is notoriously slow, requiring quadratic time in cluster size to find a cluster centre; it is also known to be highly sensitive to initialization. The resulting difficulty in clustering under rich gene-gene similarity measures has meant that existing work on such measures has not led to a widely applicable alternative to standard clustering.

We note that time delays are well accounted for in graphical model formulations (including dynamic Bayesian networks, state space models and hidden Markov models) where Markov assumptions are used to model these temporal effects. However, these approaches are computationally demanding and statistically challenging for high-dimensional data, and have for these reasons not usually been exploited to provide practical alternatives to clustering for exploratory analysis. Hierarchical clustering (see e.g. Hastie *et al*., 2001) and spectral clustering (Shi & Malik, 2000; Ng *et al*., 2002) address the related but quite distinct problem of partitioning a dataset by recursively comparing pairs of observations. In particular, these methods do not ensure that all points within a cluster are similar to a cluster mean or centre and indeed quite often make splits which lead to clusters which do not have this property.

Here, we address these open issues by putting forward an approach for finding gene modules which incorporates these key temporal features — time lags, transient correlation and inversions — but is computationally efficient enough to provide a practical alternative to standard clustering. We do so by exploiting a recently proposed message-passing-based algorithm called Affinity Propagation (AP) (Frey & Dueck, 2007) which we show, using biological data, to be robust and efficient in this setting. As a similarity measure we choose a dynamic programming formulation due to Qian *et al*. (2001); this is fast but approximate, and we confirm empirically that it is sufficiently powerful to give good results in this setting.

Our work adds to the existing literature in two main ways. First, we put forward an approach for clustering microarray time series data which captures rich temporal features yet is robust, requires little or no user input and is fast enough for routine use in microarray data analysis. For example, in an analysis of real microarray data, this finds a substantially better value of the same objective function than any of 400 runs of K-centres, while requiring a fraction of the total compute time, and no user input whatsoever. Second, we show extensive results on experimental data, highlighting the biological relevance of richer temporal features and the importance of capturing such features during clustering. We are able to cluster together members of a recently identified gene regulatory network whose profiles would not have been clustered together by traditional clustering techniques. We also find several modules which suggest hypotheses to test experimentally.

The remainder of the paper is organized as follows. We begin by reviewing basic ideas and notation for clustering and then describe the methods used here. We show results on a validated set of TF-target pairs in yeast, and on experimental data from a study of *Botrytis cinerea* infection in *Arabidopsis thaliana*. We conclude with a discussion of the shortcomings of our work, possible extensions and its relationship to other methods.

2 BACKGROUND

2.1 Notation

Let X_{it} be the mRNA expression value of gene i at time t. A time series microarray dataset, $\mathbf{X}$, is a matrix containing the expression values of genes $i \in \mathcal{I} = \{1, 2, \cdots, g\}$, for time points $t \in \mathcal{T} = \{1, 2, \cdots, T\}$. The complete expression profile for gene i is denoted $X_{i\cdot} = [X_{i1}, X_{i2}, \cdots, X_{iT}]^{\mathrm{T}}$.

2.2 Clustering

Clustering is a form of unsupervised machine learning in which observations are partitioned into groups, called clusters, such that within-cluster similarity is large relative to between-cluster

similarity. In the present setting, observations correspond to gene expression profiles $X_{i\cdot}$.

2.2.1 K-means Given a user-set number of clusters K, (Euclidean) K-means seeks to find cluster assignments $c(i)$, $c : \mathcal{I} \mapsto \mathcal{K} = \{1 \dots K\}$ and corresponding cluster means $\{\mu_k\}_{k\in\mathcal{K}}$ which minimize the following cost function:

$$J(\{c(i)\}, \{\mu_k\}) \quad = \quad \sum_{k\in\mathcal{K}} \sum_{i:c(i)=k} \|X_{i\cdot} - \mu_k\|^2 \qquad (1)$$

where, $\|\cdot\|^2$ denotes (squared) Euclidean distance and $\{c(i)\}$ and $\{\mu_k\}$ are cluster assignments and cluster means respectively.

K-means minimizes this cost function by means of an iterative procedure in which the computation of cluster means alternates with cluster assignment. Mixture-model-based approaches can be viewed as a probabilistic generalization of K-means, in which observations are assigned to clusters in a "soft" manner, under a probability model in which cluster membership is treated as a latent variable. Model fitting is usually accomplished using the EM algorithm; as is well-known, K-means itself arises as a certain limiting case of EM applied to a Gaussian mixture model.

2.2.2 K-centres Cost function Eq. (1) directly uses cluster means $\{\mu_k\}$. In contrast, a matrix of simlarities $\psi(i,j)$, $i,j \in \mathcal{I}$ between observations may not give an analogue to cluster mean. In this setting, a standard approach is to characterize a cluster by means of an observation within that cluster, referred to as the *centre* of the cluster. This formulation yields the following cost function:

$$J(\{e(i)\}) \quad = \quad - \sum_{i\in\mathcal{I} s.t. i\neq e(i)} \psi(i, e(i)) \qquad (2)$$

where, $e : \mathcal{I} \mapsto \mathcal{E} \subset \mathcal{I}, |\mathcal{E}| = K$

is a cluster assignment function which in this case maps observations to the (index of) the corresponding cluster centre.

The K-centres algorithm is a K-means-like heuristic method for optimizing Eq. (2), in which a cluster characterization step is alternated with a cluster assignment step. Absent any notion of mean, the cluster characterization step involves searching over all members of each cluster to minimize within-cluster distance; this requires quadratic time in cluster size. Moreover, K-centres must be initialized, and the initialization can affect which local maximum the method will find.

Thus, while Eq. (2) provides a natural cost function for clustering under a similarity matrix ψ, it can be difficult to obtain good clusters in practice, and moreover to do so robustly and rapidly in applications with a large number of objects to be clustered.

3 METHODS

Here we describe the methods used in the remainder of the paper. We first discuss clustering by Affinity Propagation (AP) and then the similarity measure used here.

3.1 Affinity propagation

Affinity propagation (AP) is an algorithm by which to learn cluster assignments and cluster centres under the K-centres cost function Eq. (2). Like K-centres, AP uses observations themselves to characterize clusters; however, unlike K-centres AP simultaneously considers *all* observations as candidate centres. Naïvely, this would be computationally intractable; in AP this is accomplished by an efficient message passing formulation (which can be derived as an instance of the max-sum algorithm for factor graphs). Two different kinds of messages are exchanged between observations: *responsibility* $r(i,j)$, which reflects point j's suitability as a centre for point i and *availability* $a(i,j)$, which reflects evidence in favour of i choosing j as its centre. Here we briefly describe the AP algorithm, as it used in the present application; for further details we refer the interested reader to Frey & Dueck (2007).

Update equations. AP is provided with a similarity matrix ψ^*, such as the one introduced in section 3.2.

Initially, availabilities $a(i,j)$ are set to zero; "self-similarities" $\psi^*(i,i)$ are given a user-set value s, this is discussed below. Then, responsibilities and availabilities are updated sequentially using the following:

$$r(i,j) \leftarrow \psi^*(i,j) - \max_{j':j'\neq j} \{a(i,j') + \psi^*(i,j')\} \qquad (3)$$

$$\forall i \neq j, \quad a(i,j) \leftarrow \min \left\{ 0, r(j,j) + \sum_{i':i'\notin\{i,j\}} \max\{0, r(i',j)\} \right\} \qquad (4)$$

$$a(j,j) \leftarrow \sum_{i':i'\neq j} \max\{0, r(i',j)\} \qquad (5)$$

A damping factor $\lambda \in [0,1]$ is used to prevent numerical oscillations: each message is set to a weighted combination of its value from the previous iteration and its updated value, weighted by λ and $1-\lambda$ respectively. In all our experiments we use a default value of $\lambda = 0.9$. Update equations are iterated until cluster centres remain unchanged for a user-set number of iterations (see below). Then, cluster centres $e(i)$ are given by maximizing over the sum of responsibility and availability:

$$e(i) \quad = \quad \operatorname*{argmax}_{j\in\mathcal{I}} a(i,j) + r(i,j) \qquad (6)$$

If $e(i) = i$, i itself is a cluster centre.

Algorithm parameters. The self-similarity value s influences the number of clusters discovered, higher values giving a greater number of clusters. However, in contrast to the parameter K in K-means and K-centres, this is not a hard specification; rather, the number of clusters found emerges from data, but is influenced by self-similarity s. In this sense, self-similarity is closer in spirit to a shrinkage/regularization strength or Bayesian hyper-parameter than a pre-specified number of clusters. Importantly, this means that a default value for s can give good results for a wide range of problems; in all our experiments, we set s to the median of the (off-diagonal entries of) similarity matrix ψ^*. Finally, we call convergence if cluster centres remain unchanged for 100 iterations and further set the overall maximum number of iterations to 1000.

3.2 Similarity measure

As noted in the introduction, there are now a number of biologically plausible similarity measures for gene expression time series in the literature. We choose a similarity score due to Qian *et al.* (2001) which uses alignment to find time lags in gene expression time series, as outlined below. Although approximate, this approach is both efficient and rich enough to capture not only time lags but also inversions and transient correlations, and is therefore well suited to our goals.

Given time series data X_{it} for genes $i \in \mathcal{I}$ at times $t \in \mathcal{T}$, Algorithm 1 returns a matrix $\psi(i,j)$ of similarity scores for all gene pairs (i,j). Data $X_{i\cdot}$ for each gene profile are assumed to be normalized to mean zero and standard deviation one. For a given pair (i,j) dynamic programming is used to build up a matrix Ω^+, which compares and scores each alignment between profiles $X_{i\cdot}$ and $X_{j\cdot}$. Inversion or negative co-expression is captured in a second

matrix Ω^-, whose entries are obtained in a similar manner. Finally, transient correlations are captured by explicitly forcing each entry of Ω^+ and Ω^- to be non-negative. Then, similarity score ψ is simply the highest entry in Ω^+ or Ω^-. The alignment matrices Ω^+ or Ω^- further yield a "match type", which may be positive/negative and simultaneous/delayed and describes the characteristics of the highest scoring alignment. Specifically, if $\omega^+ = \psi$ the profiles have a positive local correlation, whereas if $\omega^- = \psi$ the profiles have a negative local correlation. Likewise, if ψ is achieved at $\Omega^+_{t_1t_2}$ or $\Omega^-_{t_1t_2}$ with $t_1 = t_2$ then the local correlation is simultaneous, otherwise it is time delayed.

For AP a similarity matrix, where identical profiles have a score of zero, is constructed using the following transformation:

$$\psi^*(i,j) = \psi(i,j) - T + 1 \tag{7}$$

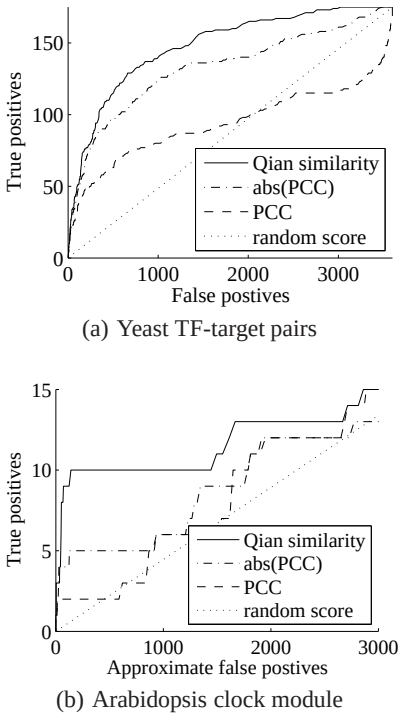
Algorithm 1 Computation of similarity measure ψ, following Qian *et al.* (2001).

(1) Initialise Ω^+_{t0}, Ω^+_{0t}, Ω^-_{t0} and Ω^-_{0t} equal to zero $\forall t \in \mathcal{T} \cup 0$.
(2) Initialise $t_1 = t_2 = 1$.
(3) Calculate $\Omega^+_{t_1t_2}$ and $\Omega^-_{t_1t_2}$:

$$\Omega^+_{t_1t_2} = \max(\Omega^+_{t_1-1t_2-1} + X_{it_1}X_{jt_2}, 0) \tag{8}$$

$$\Omega^-_{t_1t_2} = \max(\Omega^-_{t_1-1t_2-1} - X_{it_1}X_{jt_2}, 0) \tag{9}$$

(4) If $t_1 < T$ and $t_2 \leq T$ then set $t_1 = t_1 + 1$ and go to step 3.
(5) If $t_1 = T$ and $t_2 < T$ then set $t_1 = 1$ and $t_2 = t_2 + 1$ and go to step 3.
(6) Let $\omega^+ = \max_{t_1t_2}\{\Omega^+_{t_1t_2}\}$ and $\omega^- = \max_{t_1t_2}\{\Omega^-_{t_1t_2}\}$. Set: $\psi(i,j) = \max\{\omega^+, \omega^-\}$.



(a) Yeast TF-target pairs

(b) Arabidopsis clock module

Fig. 1. Validation results using microarray data. (a) ROC plots obtained from microarray data for validated examples of TF-target pairs in yeast (data from Spellman *et al.* (1998); Gasch *et al.* (2000)). Similarity score ψ outperforms both Pearson's correlation coefficient (PCC) and its absolute value. The dotted line corresponds to random guesswork. (b) ROC plots obtained from microarray data, comparing the expression profiles of genes from the *A. thaliana* circadian clock with that of random genes. Similarity score ψ outperforms the other measures of similarity, performing roughly twice as well as measures neglecting time lags.

4 RESULTS

We first show results in which we investigate whether richer temporal features are indeed useful in uncovering biological relationships. We then compare the ability of K-centres and AP to cluster real microarray data under similarity matrix ψ. Finally, we present an analysis, using our temporal clustering approach, of a microarray time course experiment we carried out to better understand the response of *A. thaliana* to infection by the pathogen *B. cinerea* (Denby, manuscript in preparation).

4.1 Validation of similarity measure ψ

We sought to investigate whether the similarity measure ψ does indeed capture biologically important relationships. To this end we used two biological examples, from yeast and Arabidopsis respectively, in which the underlying biology is relatively well understood.

TF-target pairs in yeast. The yeast genome has been well studied and provides a number of validated TF-target pairs. This makes yeast TF-target pairs well suited to a validation study. Here, we used published microarray data (Spellman *et al.*, 1998; Gasch *et al.*, 2000; Qian *et al.*, 2003) of such regulatory pairs, consisting of validated positive and negative examples. The positive examples were chosen from TRANSFAC and SCPD; negative examples were identified by finding genes without the known binding site of the transcription factor or permuting the gene (but not the transcription factors) expression profile. The expression profiles cover a total of 79 time points, which gives a relatively high time resolution in line with the general motivation for our approach. We assessed the ability of the similarity score ψ to capture underlying biology by means of a Receiver Operator Characteristic (ROC) analysis. Similarity scores $\psi(i,j)$, for each TF-target pair (positive and negative), were thresholded to yield predictions of TF-target pairs. The predictions were then compared with the list of known positive and negative pairs to yield true positive and false positive rates as a function of threshold level. Varying the threshold gives a curve which is referred to as a ROC curve; this shows the sensitivity and specificity of the analysis across all possible thresholds on a single plot, giving a comprehensive view of the ability of the score to distinguish positive and negative examples. Fig 1(a) shows ROC curves obtained from these yeast data for the similarity score ψ, the widely-used Pearson's correlation coefficient (PCC) and absolute PCC. The (expected) curve which would be obtained by random chance is also shown for comparison. Similarity score ψ performs better than both PCC and the absolute value of PCC in this instance, suggesting that the score is indeed able to detect instances of direct regulation.

Arabidopsis clock module. The results presented above pertain to direct regulatory relationships between TFs and validated targets. However, the complete set of pairwise relationships in a gene regulatory module naturally includes indirect as well as direct influences; e.g. within a module, if TF A has as its target gene B, which in turn has target C, the pair (A, C) is an example of an indirect relationship. We therefore sought to complement results from yeast TF-target pairs with a study of a well-studied gene regulatory network in *A. thaliana*. A small network of just six genes has been shown to jointly control the circadian clock in *A. thaliana* (Locke *et al*., 2006). Microarray data for these six genes were supplemented with data for a further 560 genes, chosen at random from the *A. thaliana* genome. None of the 560 genes were annotated as belonging to the circadian clock (Swarbreck *et al*., 2008). In the resulting set of pairs, those including only members of the known circadian clock module were treated as positive examples, while those with only one member of the circadian clock were considered to be false positives[1]. As the similarity measure is symmetric, we have $\binom{6}{2} \times \frac{1}{2} = 15$ positive examples and $6 \times 560 = 3,360$ negative examples. Data were obtained from leaf samples taken every 2 hours for 48 hours. ROC curves were constructed in a similar manner to the TF-target case above.

Fig. 1(b) shows ROC curves obtained in this way: similarity score ψ very clearly outperforms PCC and its absolute value in this instance. For example 10 (out of 15) true positives are obtained at a cost of 141 false positives; in comparison, PCC requires 1649 and absolute PCC requires 1783 false positives. This suggests that ψ is indeed able to detect both direct and indirect regulation, even under highly sparse conditions, i.e. when true positives are scarce relative to false positives. We note also that the vast gains relative to random selection we see using all three similarity scores confirm that the data are indeed information rich.

4.2 Comparative results

The similarity measure ψ captures a quite different notion of closeness than a straightforward vector distance; we have shown biological evidence in Fig 1 above that in the context of regulatory relationships in time series data, ψ offers a superior ability to discern validated biology. Because of this underlying difference in the notion of closeness, clustering under ψ represents a fundamentally different formulation of the clustering problem than many widely-used methods (Hastie *et al*., 2001; Ghosh & Chinnaiyan, 2001; Heard *et al*., 2005; Thalamuthu *et al*., 2006). In this sense, our approach and these widely used methods address different questions, which makes them difficult to compare directly. However, K-centres (Hastie *et al*., 2001) represents a natural choice for clustering under the similarity measure ψ; indeed, it has been used for this purpose in previous work (Qian *et al*., 2001). We therefore compared our AP-based approach with K-centres, to investigate its ability to find clusters under similarity measure ψ. We used two microarray time series; 4,489 genes over 18 time points from a published study in yeast (Spellman *et al*., 1998) and 6,000 genes over 24 time points from a study we have carried out on *A. thaliana* leaves during infection by the necrotrophic fungal pathogen *B. cinerea*.

For each dataset we applied both methods to the full set of genes and also used smaller, randomly selected subsets, to investigate dependence on dimensionality. For each regime of dimensionality, 10 runs of K-centres and one run of AP (which is deterministic) was applied to the data. Since we use the same similarity measure in both cases, the underlying cost function Eq. (2) is identical. AP was applied using default parameters; AP is able to automatically learn a good number of clusters (Frey & Dueck, 2007). To ensure a fair comparison, we set the number K of clusters for K-centres to equal the number of clusters discovered by AP in each case. Fig 2(a) shows results obtained using the yeast dataset of Spellman *et al*. (1998), which is a time course of expression profiles of genes from cells synchronised by the addition of alpha pheromone. The *A. thaliana* dataset contains the expression profiles of 6,000 genes shown to be differentially expressed between infected and so-called "mock infected" leaves (i.e. a control set of leaves not inoculated with *B. cinerea* spores, but otherwise kept in identical experimental conditions). Figure 2(b) shows results on the *A. thaliana* data. In each case, boxplots show values of the objective function obtained using K-centres; AP is deterministic and gives a single result in each case.

Fig 2(c) shows an analysis in which we used 400 K-centres runs on the full *A. thaliana* dataset, with each run allowed the same compute time as a single run of our method. Our method is completely deterministic, and therefore not subject to variation due to initial conditions or stochastic steps. It is clear that K-centres is performing significantly worse than our method at producing clusters to minimize cost function (2).

4.3 Temporal clustering of *A. thaliana* time series data

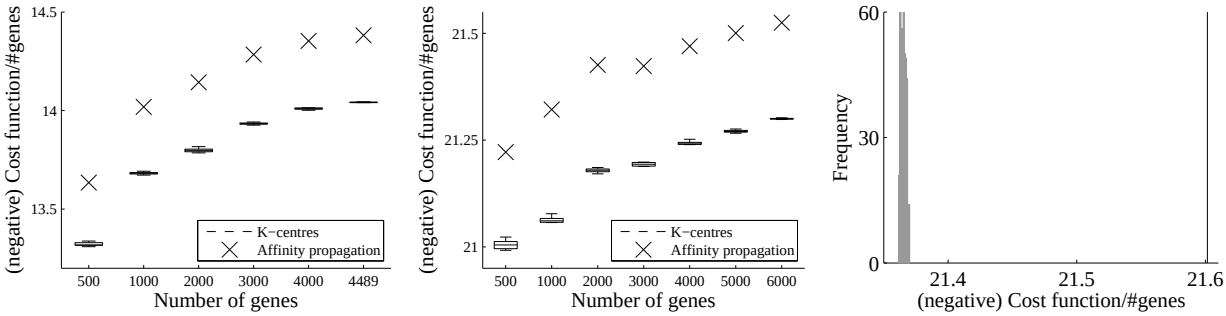
Here, we apply our method to a microarray time series dataset of gene expression in *A. thaliana* leaves during infection by the necrotrophic fungal pathogen *B. cinerea*, as described in Section 4.2. We use the VirtualPlant software platform for GO term over-representation analysis, with p-values calculated using the hypergeometric distribution (Gútierrez *et al*., 2005).

We first visually highlight the ability of our method to uncover non-obvious clusters by means of an illustrative example. Fig. 3(a) is an example of a cluster whose underlying temporal patterns are sufficiently complex as to make the cluster appear, at first glance, devoid of any coherent pattern. Fig. 3(b) shows the same cluster, adjusted for time lags and inversions: this is now highly coherent.

Application of our method produced 481 clusters; 143 of these were singleton clusters and so were ignored. In Fig. 4 we highlight several clusters which yielded modules with known interactions or novel modules which are biologically interesting.

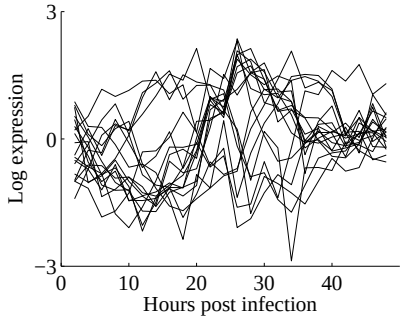
Circadian clock. Fig. 4(a) shows a cluster which appears to have a 24 hour rhythm. The cluster contains two genes encoding known components of the circadian clock module. Gene *GI* is found to score highly with *LHY* with a delayed and inverted match. The delayed and inverted relationship between the two expression profiles fits extremely well with the known role of *LHY* as a transcriptional repressor of *GI* (Locke *et al*., 2006). In addition, another member of the cluster, At1g56300, belongs to a class of genes known as Rapid Wounding Response (RWR) genes, which are known to be regulated by the circadian clock (Walley *et al*., 2007).

[1] Despite these precautions, it is possible that some of the 560 genes are circadianly regulated, as their roles may not currently be fully known. However, it is highly unlikely that any more than a small minority are so regulated.

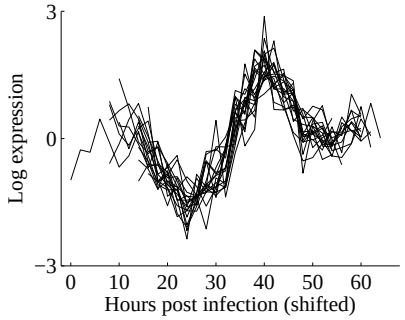


(a) AP vs. K-centres in yeast data (b) AP vs. K-centres in Arabidopsis data (c) AP vs. K-centres in Arabidopsis data, 6000 genes

Fig. 2. Here the method proposed in Qian *et al*. (2001) is compared to our method. (a) They are both applied to data from Spellman *et al*. (1998), a time series consisting of 4,489 genes over 18 time points. Various subsets of this are clustered and the cost function, as given in Eq. (2) and then divided by the number of genes in the subset, is reported. 10 runs of K-centres each allowed to take as long as a single run of AP were applied to the data. (b) Both methods are applied to data from *A. thaliana* leaves during infection by the necrotrophic fungal pathogen *B. cinerea*. Various subsets of this are clustered and the cost function, as given in Eq. (2) and then divided by the number of genes in the subset, is reported. 10 runs of K-centres each allowed to take as long as a single run of AP were applied to the data. (c) Here the *A. thaliana* data is clustered again by both methods, but with 400 runs of K-centres (shown in the grey histogram) each allowed to take as long as a single run of AP (black line, representing the result of a single run of AP).



(a) Complex temporal cluster



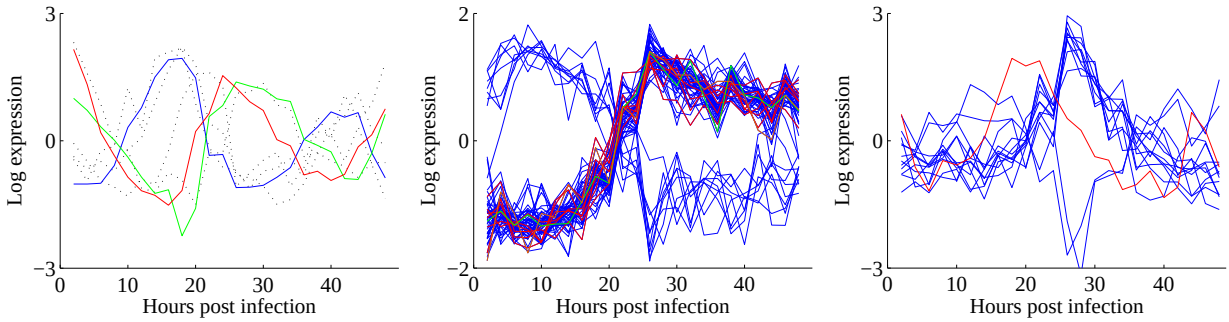
(b) Complex temporal cluster, adjusted for delay and inversions

Fig. 3. (a) A cluster returned by our method. (b) The same cluster as in the previous figure, adjusted for time delays and anti-correlation. Some profiles in this plot have been shifted in time and/or vertically inverted according to their original match type.

The *de novo* discovery of a small cluster containing these genes is striking in light of the fact that the relationship between these genes took many years and much research effort to uncover. To the best of our knowledge, the remaining cluster members have no known link to the circadian clock; however, given the highly validated nature of other cluster members, these further genes provide intriguing hypotheses for additional downstream targets.

Ethylene response. Fig. 4(b) shows a second cluster whose members form a striking and biologically coherent group. It is noteworthy that this cluster contains a regulator and known target genes of this regulator. The TF *ORA59* (At1g06160) is in this cluster, along with six genes (At1g59950, At2g43580, At3g23550, At3g56710, At4g11280, At4g24350) that have been previously found to be upregulated in an inducible overexpressor line of *ORA59* (Pré *et al*., 2008). These genes are also upregulated in the present experiment. Moreover, *ORA59* and another TF, *ERF1* (At3g23240), are believed to jointly regulate *PDF1.2* (Pré *et al*., 2008) and *ERF1* is also found in this cluster. *PDF1.2* itself is not in the dataset as there is no probe for it on the microarrays used. Both *ORA59* and *ERF1* are known to respond to the plant hormone ethylene; the cluster also has an over-representation, significant at 1%, of the GO term response to ethylene stimulus. Little is known in Arabidopsis about the relative timing of expression of TFs and their direct targets. However, in this case the time resolution of the dataset (2 hr) is apparently not sufficient to pick up a delay between the expression of the regulator *ORA59* and its targets.

Response to abscisic acid. The cluster of 13 genes shown in Fig. 4(c) highlights a novel putative transcriptional module. The only TF in this cluster, At1g71030 (*AtMYBL2*) scores highly for a match with the other genes with a time delay of 6 hours. This cluster has an over-representation, significant at 1%, of the GO term "response to abscisic acid (ABA)" and as such may represent a transcriptional module involved in signalling in response to this hormone. Intriguingly, ABA has been shown to play a role in the interaction between *B. cinerea* and plant hosts (Audenaert *et al*., 2002; AbuQamar *et al*., 2006).



(a) Circadian cluster (b) Cluster containing *ORA59* and *ERF1* (c) Cluster containing *AtMYBL2*

Fig. 4. Clusters found by applying our method to biological data, with default parameters. Data represents *Arabidopsis thaliana* gene expression levels following infection by *Botrytis cinerea*. (a) A circadian module. *LHY* (in blue) is known to be a transcriptional repressor of *GI* (in green). At1g56300 (in red) is a Rapid Wounding Response gene, which are known to be regulated by the circadian clock. Here black dotted lines represent the expression levels of four additional cluster members. (b) A cluster containing 6 genes co-regulated by *ORA59* (in red), *ORA59* (in orange) and gene *ERF1* (in green) that is believed to jointly regulate *PDF1.2* with *ORA59* (Pré *et al*., 2008). (c) A putative transcriptional module. *AtMYBL2* (in red) is the only known transcription factor in this cluster, and peaks 3 time points before the rest of the genes.

5 DISCUSSION

In this paper we have introduced a clustering methodology that can reveal relatively complex temporal features in gene expression time series datasets. Our method is complementary to standard clustering approaches, but aimed specifically at high resolution time series and regulatory modules whose expression profiles have complex temporal relations. Here we discuss the shortcomings of our method, discuss possible extensions and the relationship of our method to others.

As transcriptional assays continue to mature higher resolution datasets are becoming more common; our method is best suited to data with (relatively) high temporal resolution, e.g. more than ten time points. Time series data with fewer time points will naturally give a higher chance of spurious correlations or missed time lags.

The method used here is able to detect transient co-expression, but is not as sensitive as biclustering methods to events occurring only within short windows of time. This is due to the conservative approximation strategy of Qian *et al*. (2001), that divides the overall score by the total number of time points rather than the number of time points where co-expression occurs. We could improve this by giving each ψ value a p-value using an empirical null distribution. For example, a local correlation across 5 time points could be compared to alignments of 5 time points in random expression profiles. A matrix could then be constructed from the p-values and clustered as described above. This would aid in identifying clusters that contain genes that are transiently co-expressed.

The deterministic approach we have used for alignment is effectively a (constrained) time-warping. An interesting extension would be to carry out alignment within a probabilistic framework using a Hidden Markov Model (Rabiner, 1989; Eddy, 1998). However, in such an approach the design of the state space would be crucial in capturing realistic gene expression time series using conventional i.i.d. Gaussian observation models. Moreover the resulting computational burden for all pairwise comparisons of $\sim 10^4$ genes would be considerably greater than the method used here, which is fast enough for interactive use as an exploratory tool.

As AP is an appropriate method to cluster arbitrary matrices of similarity, it provides a flexible framework in which to carry out further work in incorporating complementary information in the similarity measure, e.g. additional time series of the same genes under different environmental conditions, the identity of TFs, presence of known TF binding sites in a gene's promoter, protein-protein interactions, etc.

A recent paper by Smith *et al*. (2009) demonstrated a method called SCOW for aligning the profile of a gene with its profile in another time series. This is subtly different from clustering the profiles of different genes in the same time series, for example, shorting is not appropriate in this case. It also allows for unequal sampling. The problem of unequal sampling was partially treated in Qian *et al*. (2001), but could certainly be improved. One way that suggests itself is to record the spacing between time points, and on the basis of that allow skips in matrices Ω^+ and Ω^- that are acceptable given the spacing.

ACKNOWLEDGEMENT

We would like to thank anonymous referees for their constructive comments. S.K. and S.Mc. are supported by an Engineering and Physical Sciences Research Council/Biotechnology and Biological Sciences Research Council grant to Warwick Systems Biology Doctoral Training Centre. This work was supported by Biotechnology and Biological Sciences Research Council [PRESTA Project, grant number BB/F005806/1] to V.B.W., J.B. and K.D. SM is partially supported by the Engineering and Physical Sciences Research Council. We would like to acknowledge Edward Morrisey for inspiration. Dedicated to Carl Blakey.

REFERENCES

AbuQamar, S. *et al*. (2006) Expression profiling and mutant analysis reveals complex regulatory networks involved in *Arabidopsis* response to *Botrytis* infection. *The Plant Journal*, **48**, 28-44.

Audenaert, K., DeMeyer, G.B., Höfte, M.M. (2002) Abscisic acid determines basal susceptibility of tomato to Botrytis cinerea and suppresses salicylic acid-dependent signaling mechanisms. *Plant Physiol.*, **128 (2)**, 491-501.

Balasubramaniyan, R., Hüllermeier, E., Weskamp, N., Kämper, J. (2004) Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics*, **21 (7)**, 1069-1077.

Bar-Joseph, Z. *et al*. (2003) Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, **22 (11)**, 1337-1342.

Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14 (9)**, 755-63.

Frey, B. and Dueck, D. (2007) Clustering by Passing Messages Between Data Points. *Science*, **315**, 972-976.

Gasch, A.P. *et al*. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell.*, **11**, 4241-4257.

Ghosh, D. and Chinnaiyan, A.M. (2007) Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, **18 (2)**, 275-286.

Gútierrez, R.A., Shasha, D.E., Coruzzi, G.M. (2005) Systems biology for the virtual plant. *Plant Physiology*, **138**, 550-554.

Hartigan, J.A. (1972) Direct clustering of a data matrix. *Journal of the American Statistical Association*, **67 (337)**, 123-129.

Hastie, T., Tibshirani, R. and Friedman, J.H. (2001) *The Elements of Statistical Learning*. Springer-Verlag.

Heard, N.A. *et al*. (2005) Bayesian coclustering of *Anopholes* gene expression time series: Study of immune defense response to multiple experimental challenges. *Proc. Nat. Acad. Sci.*, **102 (47)**, 16939-16944.

Lazzeroni, L. and Owen, A. (2002) Plaid models for gene expression data. *Statistica Sinica*, **12 (2002)**, 61-86.

Locke, J.C.W. *et al*. (2006) Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*. *Molecular Systems Biology*, **2006**, 1-6.

Madeira, S.C. and Oliveira, A.L. (2005) A linear time biclustering algorithm for time series gene expression data. *Lecture Notes In Computer Science*, **3692**, 3806-3807.

Meng, J., Gao, S.J., Huang, Y. (2009) Enrichment constrained time-dependent clustering analysis for finding meaningful temporal transcription modules. *Bioinformatics*, **25 (12)**, 1521-1527.

Ng, A., Jordan, M., Weiss, Y. (2002) On spectral clustering: analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14*. MIT Press.

Pré, M. *et al*. (2008) The AP2/ERF domain transcription factor ORA59 integrates jasmonic acid and ethylene signals in plant defense. *Plant Physiology*, **147**, 1347-1357.

Qian, J. *et al*. (2001) Beyond synnexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new biologically relevant interactions. *J. Mol. Biol.*, **314**, 1053-1056.

Qian, J., Lin, J., Luscombe N.M., Yu, H., Gerstein, M. (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, **22 (13)**, 1917-1926.

Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77, 257?286 (1989)

Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257-286 (1989).

Schmitt Jr, W.A., Raab, R.M., Stephanopoulos, G. (2004) Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. *Genome Res.*, **2004 (14)**, 1654-1663.

Segal, E. *et al*. (2003) Module networks: identifying regulatory networks and their condition specific regulators from gene expression data. *Nat. Genet.*, **34**, 166-176.

Shi, J. and Malik, J. (2000) Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8):888-905.

Smith, A.S., Vollrath, A., Bradfield, C.A., Craven, M. (2009) Clustered alignments of gene-expression data. *Bioinformatics*, **25 (12)**, 1521-1527.

Spellman, P.T.*et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.*, **9**, 3273-3297.

Swarbreck, D. *et al*. (2008) The Arabidopsis information resource (TAIR): gene structure and function annotation. *Nucleic Acids Research*, **36**, 1009-1014.

Thalamuthu, A., Mukhopadhyay, I., Zheng, X., Tseng, G.C. (2006) Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22 (19)**, 2405-2412.

Walley, J.W. *et al*. (2007) Mechanical stress induces biotic and abiotic stress responses via a novel cis-element. *PLoS Genet.*, **3 (10)**, 1800-1812.

Yona,G., Dirks, W., Rahman, S., Lin, D.M. (2006) Effective similarity measures for expression profiles. *Bioinformatics*, **22 (13)**, 1616-1622.