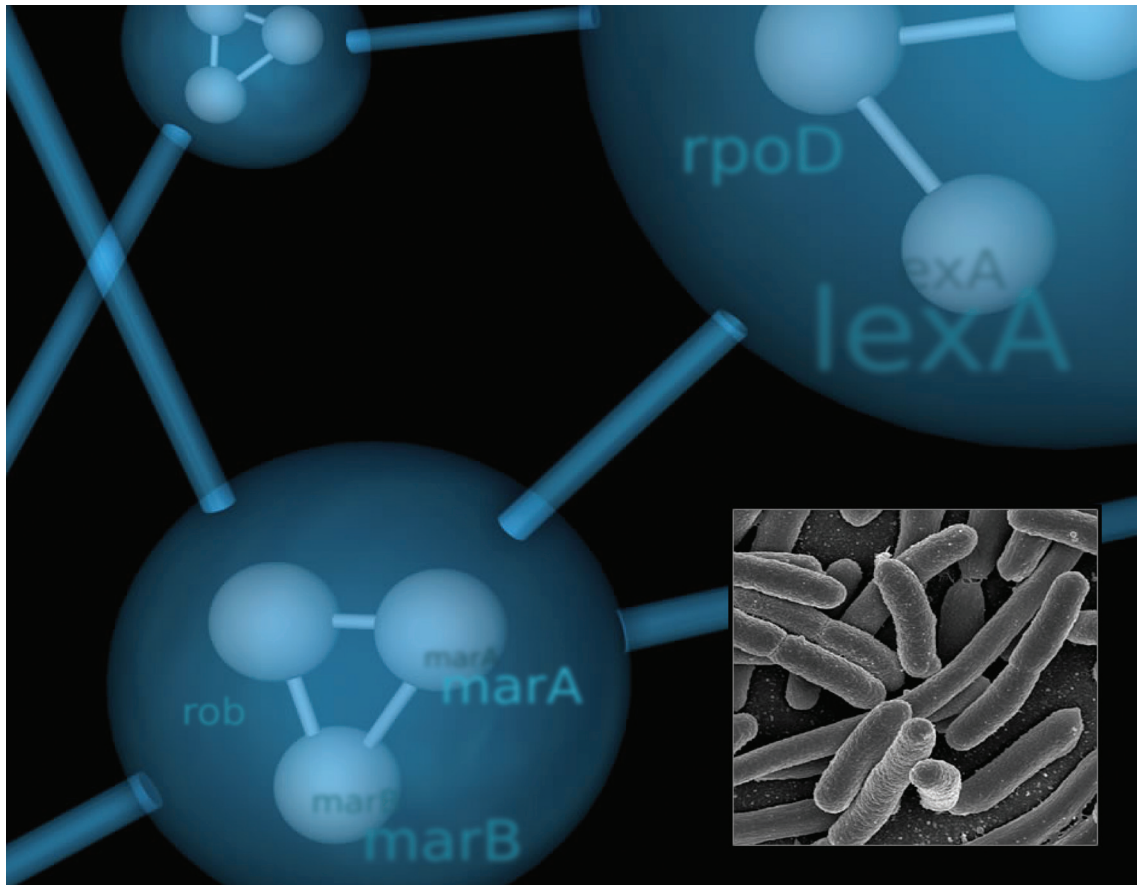


Molecular BioSystems

This article was published as part of the

Computational and Systems Biology
themed issue

Please take a look at the full [table of contents](#) to access the
other papers in this issue.



Finding explained groups of time-course gene expression profiles with predictive clustering trees†

Ivica Slavkov,^a Valentin Gjorgjioski,^a Jan Struyf^b and Sašo Džeroski^{*a}

Received 9th July 2009, Accepted 8th January 2010

First published as an Advance Article on the web 19th February 2010

DOI: 10.1039/b913690h

In biology, analyzing time course data is usually a two-step process, beginning with clustering of similar temporal profiles. After the initial clustering, depending on the expert's knowledge, descriptions of the clusters are elucidated (*e.g.*, Gene Ontology terms that are enriched in the clusters). In this paper, we investigate the application of so-called predictive clustering trees (PCTs) for the analysis of time series data. PCTs are a part of a more general framework of predictive clustering, which unifies clustering and prediction. Their advantage over usual clustering approaches is that they partition the time course data into homogeneous clusters while at the same time providing symbolic descriptions of the clusters. We evaluate our approach on multiple yeast microarray time series datasets. Each dataset records the change over time in the expression level of yeast genes as a response to a specific change in environmental conditions. We demonstrate that PCTs are able to cluster genes with similar temporal profiles, yield a predictive model of the temporal profiles of genes based on a cluster prototype, and provide cluster descriptions, all in a single step.

1. Introduction

Gene expression is a temporal process that is highly regulated. Much work in bioinformatics studies this process in order to better understand the function of individual genes and to gain insight into complete biological systems. The task most commonly addressed in this context is the task of clustering time series of gene expression data, where the aim is to discover groups of genes with similar temporal profiles of expression and to find common characteristics of the genes in each group. Clustering genes by their time expression pattern is important, because genes that are co-regulated or have a similar function will have similar temporal profiles under certain conditions.

The purpose of our research is to develop a clustering approach that is well suited for analyzing short time series, and to demonstrate its usefulness on time series expression data. Besides finding clusters, *e.g.*, groups of genes, we also aim to find descriptions/explanations for the clusters. Instead of first clustering the expression time series and elucidating the characteristics of the obtained clusters later on (as done in, *e.g.*, ref. 1), we perform so-called constrained clustering, which yields both the clusters and their symbolic descriptions all in one step.

The constrained clustering is performed by using predictive clustering trees (PCTs), which are a part of a more general framework, namely predictive clustering. This general framework

of predictive clustering combines clustering and prediction.² Predictive clustering partitions a given dataset into a set of clusters such that the instances in a given cluster are similar to each other and dissimilar to the instances in other clusters. In this sense, predictive clustering is identical to regular clustering.³ The difference is that predictive clustering associates a predictive model to each cluster. This model assigns instances to clusters and provides predictions for new instances. So far, decision trees^{2,4} and rule sets⁵ have been used in the context of predictive clustering.

This paper investigates how predictive clustering can be applied to cluster time series,⁶ *i.e.*, sequences of measurements of a continuous variable that changes over time. For example, Fig. 1a shows eight time series partitioned into three clusters: cluster C_1 contains time series that increase and subsequently decrease, C_2 has mainly decreasing time series and C_3 mainly increasing ones. Fig. 1b shows a so-called predictive clustering tree (PCT) for this set of clusters. The tree represents a hierarchical clustering of the time series, where each leaf corresponds to one of the three clusters. At each leaf, a prototype is given for the cluster. This is the predictive model associated with the cluster. Finally, each cluster is described by a set of conditions. For example, cluster C_1 includes all genes that are annotated with the Gene Ontology terms “GO:0043232” and “GO:0000313”.

We first propose an extension of the general PCT induction algorithm² to the task of time series clustering. We use the name “Clus-TS” (Clustering-Time Series) for this extension. From a computational viewpoint, applying the PCT induction algorithm to time series clustering is non-trivial because the general algorithm requires computing a centroid for each cluster and for most distance measures suitable for time series clustering, no closed algebraic form centroid is known.

^a Dept. of Knowledge Technologies, Jožef Stefan Institute, Slovenia.

E-mail: Ivica.Slavkov@ijs.si, Valentin.Gjorgjioski@ijs.si, Sašo.Džeroski@ijs.si

^b Dept. of Computer Science, Katholieke Universiteit Leuven, Belgium. E-mail: Jan.Struyf@cs.kuleuven.be

† This article is part of a Molecular BioSystems themed issue on Computational and Systems Biology.

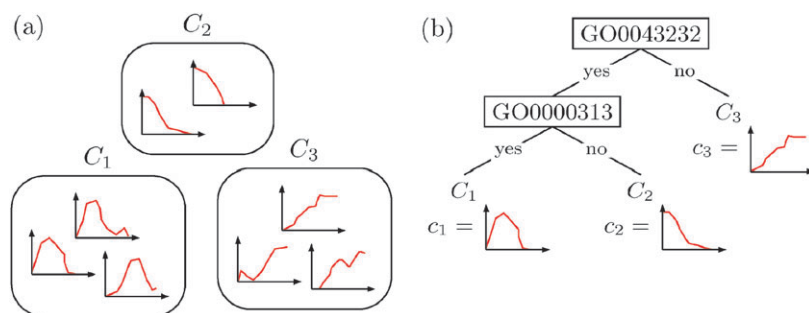


Fig. 1 (a) A set of time series clustered into three clusters. (b) A predictive clustering tree associated with this clustering. Each leaf of the tree corresponds to one cluster and stores the cluster's prototype which is used for prediction.

We also demonstrate the usefulness of Clus-TS on several time series datasets generated by microarray expression profiling.⁷ Each dataset records the change over time in the expression level of yeast genes in response to a different type of change in environmental conditions. There has been significant research related to clustering this type of short time series gene expression data,^{1,8–13} using several different distance measures. Our approach uses an alternative distance measure (that mainly takes the shape of the time series into account) and constructs clusters together with their explanations in terms of a given set of descriptive features. Here, as descriptive features, we consider terms from the Gene Ontology (GO),¹⁴ but this can be extended to any other type of gene descriptions (*e.g.*, KEGG pathways,¹⁵ regulatory motifs). The GO terms appear in the internal nodes of the PCT (Fig. 1b) and provide a symbolic description of the clusters.

In the remainder of the paper, we first give an overview of related work. We next present our methodology in more detail: this includes a description of the predictive clustering framework, the PCT induction algorithm, the distance measure used for clustering, and the methodology for evaluating predictive error. We then present the results of our analysis of the yeast gene expression time profiles, where we evaluate our approach in terms of predictive error and the usefulness of the descriptions derived from the PCTs. We conclude the paper with a discussion in light of the presented results and methodology.

2. Related work

A large body of work has been devoted to the task of analyzing expression time series data. Bar-Joseph¹⁶ presents an overview of the most important aspects that are relevant when analyzing expression time series data. This includes experimental design, data preprocessing (dealing with differences in sampling rates, missing values, and noise), finding significant genes, modeling gene interaction, and clustering expression time series.

Many different clustering algorithms³ have been used to cluster expression time series data. The most well-known algorithm is probably UPGMA, which was proposed by Eisen *et al.* in 1998¹⁷ and performs hierarchical clustering based on correlation. More recently, several advanced time series clustering methods have been presented. These model the time series, for example, using spline curves,^{8,10} an autoregressive model,^{11,13} or a mixture of hidden Markov models.¹²

Datta and Datta⁹ compare six clustering algorithms for expression time series data experimentally. Their comparison includes two hierarchical clustering algorithms (among which UPGMA), divisive clustering (Diana), fuzzy clustering (Fanny), a model based clustering method, and *k*-means. They found Diana to be a solid and robust performer across different evaluation measures. A review of the most common evaluation measures for clustering is provided by Handl *et al.*¹⁸

Often, the clustering methods are not applied to all genes, but only to genes that do respond to the change in environmental conditions or treatment. A gene responds to the treatment if the null hypothesis stating that its expression over time is constant can be rejected.¹⁹ Several methods have been proposed to identify such genes and a comparison can be found in Mutarelli *et al.*²⁰ An advantage of our method is that it detects such genes during the clustering process itself by assigning confidence values to genes.

Due to the cost of microarray analysis and of obtaining samples, most expression time series are relatively short (≤ 8 points). Ernst *et al.*¹ propose a clustering method designed for such short time series. Their method creates all possible expression profiles under the constraint that the maximal expression change between subsequent time points is bounded by a fixed number of units. It then assigns time series to the closest profile (in terms of correlation) thereby forming clusters. Our method is also tailored to short time series, but instead of using correlation, we opt for a qualitative distance measure that can be reliably estimated from short time series.

After clustering, Ernst *et al.*¹ label the clusters by finding GO categories that are significantly enriched in the clusters. Our method also provides a description for each cluster in terms of GO categories, but finds these during the constrained clustering process itself. As a result, all genes in a cluster are guaranteed to belong to the GO categories from the description. This is closely related to the constrained clustering method by Sese *et al.*²¹ The main difference is that their method deals with static gene expression data and not with time series, and that their cluster descriptions are restricted to item-sets.

3. Methodology

3.1 Prediction, clustering, and predictive clustering trees

Predictive modeling aims at constructing models that can predict a target property of an object from a description of

the object. Predictive models are learned from sets of examples, where each example has the form (D, T) , with D being an object description and T a target property value. For example, D can be the measured gene expression levels of a certain sample, and T whether the corresponding tissue is cancerous or healthy. While a variety of representations, ranging from propositional to first order logic, have been used for D , T is almost always a single target attribute called the class, which is discrete for classification problems or continuous for regression problems.

Clustering,³ on the other hand, is concerned with grouping objects into subsets of objects (called clusters) that are similar with respect to their description D : this is called distance based clustering. There is no target property defined in clustering tasks. In conventional clustering, the notion of a distance (or conversely, similarity) is crucial: examples are considered to be points in a metric space and clusters are constructed such that examples in the same cluster are close according to a particular distance defined on the descriptive space D . A centroid (or prototypical example) may be used as a representative for a cluster. The centroid is the point with the lowest average (squared) distance to all the examples in the cluster, *i.e.*, the mean or medoid of the examples. Hierarchical clustering and k -means clustering are the most commonly used algorithms for this type of clustering.³

Predictive clustering² combines elements from both prediction and clustering. As in clustering, we seek clusters of examples that are similar to each other. The distance measure is defined on $D \cup T$, taking both the descriptive part and the target property into account. In addition, a predictive model must be associated to each cluster. The predictive model assigns new instances to clusters based on their description D and provides a prediction for the target property T . A well-known type of model that can be used to this end is the decision tree.²² A decision tree that is used for predictive clustering is called a predictive clustering tree (PCT, Fig. 1b). Each node of a PCT represents a cluster. The conjunction of conditions on the path from the root to that node gives a description of the cluster. Essentially, each cluster has a symbolic description in the form of a rule (IF conjunction of conditions THEN cluster)[†], while a tree structure represents the hierarchy of clusters. Clusters that are not on the same branch of a tree do not overlap.

In Fig. 1, the description D of a gene consists of GO terms with which the gene is annotated, and the target property T is the time course expression recorded for that gene. In general, we could include both D and T in the distance measure. We are, however, most interested in the time course part. Therefore, we define the distance measure only on T . We consider the so-called qualitative distance measure (QDM),²⁴ described in section 3.5. The resulting PCT (Fig. 1b) represents a clustering that is homogeneous w.r.t. T and the internal nodes of the tree provide a symbolic description of the clusters. Note that a PCT can also be used for prediction: we can use the tree to assign a new instance to a leaf and take the centroid (denoted with c_i in Fig. 1b) of the corresponding cluster as a prediction.

[†] This idea was first used in conceptual clustering.²³

3.2 Building predictive clustering trees

The generic algorithm for constructing PCTs² is presented in Table 1. It is a variant of the standard greedy recursive top-down decision tree induction algorithm used in ref. 22. It takes as input a set of instances I ; in our case these are genes described by GO terms and their associated time course measurements. The algorithm calls the procedure **BestTest** (Table 1, right) to search for the best acceptable test (GO term) that can be put in a node. If such a test t^* can be found then the algorithm creates a new internal node labeled t^* , splits the instances into several subsets (partition P^*) according to the outcome of the test for each instance, and calls itself recursively to construct a tree for each of the subsets in P^* . If no acceptable test can be found, then the algorithm creates a leaf, and the recursion terminates. The procedure “Acceptable” defines the stopping criterion of the algorithm, *e.g.*, specifying maximum tree depth or a minimum number of instances in each leaf. We enforce different constraints on the size of the tree by means of the post pruning method proposed by Garofalakis *et al.*,²⁵ which employs dynamic programming to find the most accurate subtree no larger than a given number of leaves.

Up till here, the algorithm is identical to a standard decision tree learner. The main difference is in the heuristic that is used for selecting the tests. For PCTs, this heuristic is the reduction in variance (weighted by cluster size, see line 6 of **BestTest**). Maximizing variance reduction maximizes cluster homogeneity. The next section discusses how cluster variance can be defined for time series.

An implementation of the PCT induction algorithm is available in the Clus system, which can be obtained at <http://www.cs.kuleuven.be/~dtai/clus>.

3.3 Computing cluster variance

The PCT induction algorithm requires a measure of cluster variance in its heuristics. The variance of a cluster C can be defined based on a distance measure as

$$Var(C) = \frac{1}{|C|} \sum_{x \in C} d^2(X, c), \quad (1)$$

Table 1 Pseudo-code for the algorithm Clus that induces predictive clustering trees (PCTs). The two key subroutines of the algorithm are **BestTest**(I) and **Centroid**(I). The first selects the best test t^* among the possible tests, according to the heuristic h , which for each test t measures the reduction of variance between the dataset I and the partition $P = I_1, I_2$ produced by the test. The second procedure calculates the cluster centroid

procedure PCT(I)	procedure BestTest(I)
1: $t^* = \text{BestTest}(I)$	1: $(t^*, h^*, P^*) = (\text{none}, 0, \emptyset)$
2: if $t^* \neq \text{none}$ then	2: for each possible test t do
3: for each $I_k \in P^*$ do	3: $I_1 = \{e \in I \mid t(e) = \text{true}\}$
4: $tree_k = \text{PCT}(I_k)$	4: $I_2 = I \setminus I_1$
5: return $\text{node}(t^*, \bigcup_k \{tree_k\})$	5: $P = \{I_1, I_2\}$
6: else	6: $h = Var(I) - \sum_{I_k \in P} \frac{ I_k }{ I } Var(I_k)$
7: return $\text{leaf}(\text{Centroid}(I))$	7: if $(h > h^*) \wedge \text{Acceptable}(t, P)$
procedure Centroid(I)	then $(t^*, h^*, P^*) = (t, h, P)$
return $\text{argmin}_q \sum_{x \in I} d^2(x, q)$	8: return t^*

with c the cluster centroid of C . To cluster time series, d should be a distance measure defined on time series, such as the QDM defined in section 3.5.

The centroid c can be computed as $\operatorname{argmin}_q \sum_{X \in C} d^2(X, q)$. We consider two possible representations for c : (a) the centroid is an arbitrary time series, and (b) the centroid is one of the time series from the cluster (the cluster prototype). In representation (b), the centroid can be computed with $|C|^2$ distance computations by substituting q with each time series in the cluster. In representation (a), the space of candidate centroids is infinite. This means that either a closed algebraic form for the centroid is required or that one should resort to approximative algorithms to compute the centroid. No closed form for the centroid is known in representation (a) for the QDM distance.

An alternative way to define cluster variance is based on the sum of the squared pairwise distances (SSPD) between the cluster elements, *i.e.*,[§]

$$\operatorname{Var}(C) = \frac{1}{2|C|^2} \sum_{X \in C} \sum_{Y \in C} d^2(X, Y). \quad (2)$$

The advantage of this approach is that no centroid is required. It also requires $|C|^2$ distance computations. This is the same time complexity as the approach with the centroid in representation (b). Hence, using the definition based on a centroid is only more efficient if the centroid can be computed in time linear in the cluster size. This is the case for the Euclidean distance in combination with using the pointwise average of the time series as centroid. For QDM no such centroids are known. Therefore, we choose to estimate cluster variance using the SSPD.

A second advantage is that (2) can be easily approximated by means of sampling, *e.g.*, by using,

$$\operatorname{Var}(C) = \frac{1}{2|C|m} \sum_{X \in C} \left(\sum_{Y \in \operatorname{sample}(C, m)} d^2(X, Y) \right), \quad (3)$$

with $\operatorname{sample}(C, m)$ a random sample without replacement of m elements from C , instead of (2) if $|C| \geq m$. The computational cost of (3) grows only linearly with the cluster size. In the experimental evaluation, we only use (3), as a previous experimental comparison shows only small differences between (2) and (3) (results not shown).

The PCT induction algorithm places cluster centroids in its leaves, which can be inspected by the domain expert and used as a prediction. For these centroids, we use representation (b) as discussed above.

3.4 Estimating the predictive error of PCTs

PCTs make predictions just like regular decision trees.²² They sort each test instance into a leaf and assign as prediction the label of that leaf. PCTs label their leaves with the training set centroids of the corresponding clusters.

To evaluate the predictive performance of PCTs, we first need an error measure and also a method to estimate it. For an

error measure we use the root mean squared error (RMSE), which is defined as:

$$\operatorname{RMSE}(I, T) = \sqrt{\frac{1}{|I|} \sum_{X \in I} d^2(T(X), \operatorname{series}(X))}, \quad (4)$$

with I the set of test instances, T the PCT that is being tested, $T(X)$ the time series predicted by T for instance X , $\operatorname{series}(X)$ the actual series of X , and d the qualitative time course distance measure (described in section 3.5).

For estimating the predictive performance of the PCTs we use k fold cross-validation. In cross-validation the dataset D is first split into k random subsets $\{D_1, D_2, \dots, D_k\}$. We then use $k - 1$ subsets to build the predictive model (in this case the PCT) and we record its error (*i.e.*, RMSE) on the left-out subset(fold). We repeat this k times, each time leaving out a different subset for testing the error. We obtain the final error estimate by averaging the errors obtained for all of the n instances of the dataset D .

$$\operatorname{err} = \frac{1}{n} \sum_{i \in D} \operatorname{err}(\operatorname{PCT}(D_{-i}), D_i) \quad (5)$$

3.5 Qualitative distance measure

Several distance measures have been defined for time series. If all time series have the same length then one can represent them as real valued vectors and use standard vector distance measures such as the Euclidean or Manhattan distance. It is also possible to use a correlation based measure to determine the degree of linear dependence between two time-series.¹⁷ Dynamic Time Warping (DTW)²⁶ is appropriate to capture non-linear distortion along the time axis and it is suitable if the time series are not properly synchronized (this is useful if one is delayed, or if the two time series are not of the same length).

These measures are, however, not always appropriate for time course clustering, and in particular not for analyzing the short time courses of expression data. The simple Euclidean or the DTW distance mainly capture the difference in scale and baseline. If a given time series is identical to a second time series, but scaled by a certain factor or offset by some constant, then the two time series will be distant (Fig. 2). Correlation is difficult to properly estimate if the number of observations is small (*i.e.*, short time course data) and it only captures the linear dependencies between the time series.

For our application (*i.e.*, clustering short time course gene expression data), the differences in scale and size are not of great importance; only the shape of the time series matters. Namely, we are interested in grouping together time-course profiles of genes that react in the same way to a given condition, regardless of the intensity of the up- or down-regulation.

For that reason, we use the qualitative distance measure proposed by Todorovski *et al.*²⁴ It is based on a qualitative comparison of the shape of the time series. Consider two time series X and Y (Fig. 2). Then choose a pair of time points i and j and observe the qualitative change in the value of X and Y at these points. There are three possibilities: increase ($X_i > X_j$), no-change ($X_i \approx X_j$), and decrease ($X_i < X_j$). d_{qual} is obtained

[§] The factor 2 in the denominator of (2) ensures that (2) is identical to (1) for the Euclidean distance.

by summing the difference in qualitative change observed for X and Y for all pairs of time points, *i.e.*,

$$d_{\text{qual}}(X, Y) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{2 \cdot \text{Diff}(q(X_i, X_j), q(Y_i, Y_j))}{N \cdot (N - 1)}, \quad (6)$$

with $\text{Diff}(q_1, q_2)$ a function that defines the difference between different qualitative changes (Table 2, Fig. 2). Roughly speaking, d_{qual} counts the number of disagreements in change of X and Y .

QDM does not have the drawbacks of correlation based measures. First, it can be computed for very short time series, without decreasing the quality of the estimate. Second, it captures the similarity in shape of the time series, regardless of whether their dependence is linear or non-linear (Fig. 2).

Table 2 The definition of $\text{Diff}(q_1, q_2)$

$\text{Diff}(q_1, q_2)$	Increase	No-change	Decrease
Increase	0	0.5	1
No-change	0.5	0	0.5
Decrease	1	0.5	0

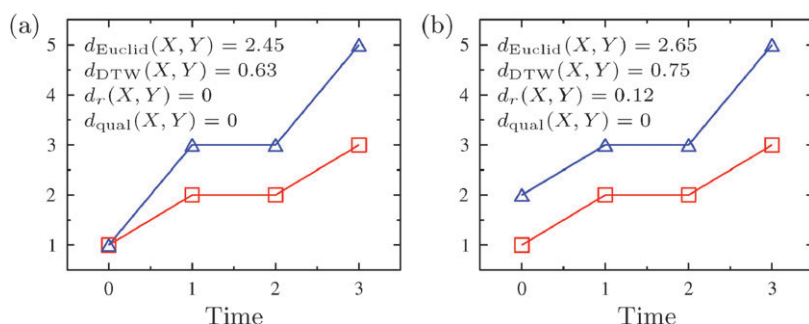


Fig. 2 Comparison of four distance measures for time series. Time series (a) are linearly related resulting in $d_r(X, Y) = 0$. Time series (b) are non-linearly related, but still have a similar shape, resulting in $d_{\text{qual}}(X, Y) = 0$.

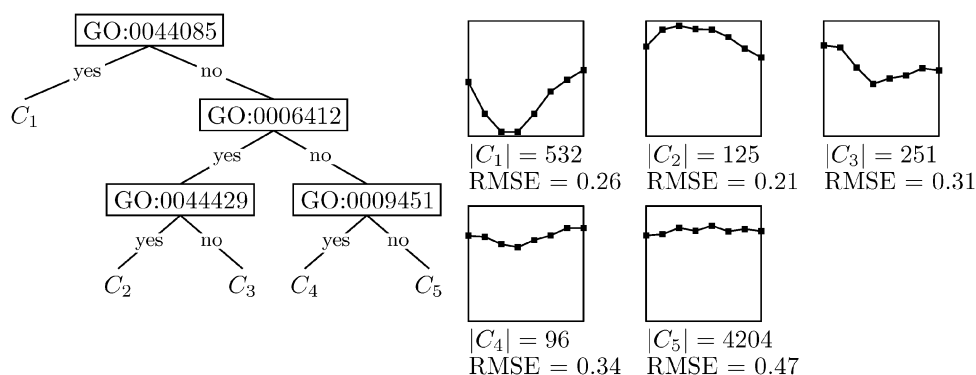


Fig. 3 On the left-hand side, we show a sample PCT with 5 leaves, produced for the diamide treatment dataset. The GO terms that appear in the nodes are used as descriptions for clusters C_1 to C_5 , found at the leaves of the tree. On the right-hand side, we show each predicted cluster prototype, and its related cluster size and RMSE. Clusters C_1 to C_3 show significant temporal changes in gene expression and have a relatively low error. Cluster C_1 includes genes that have an immediate and very significant down-regulation during diamide exposure. C_3 shows the same tendency, except the genes are less down-regulated and there is a short time-lag in their response. Cluster C_2 contains genes that are up-regulated during stress. All three cluster prototypes show that changes in gene expression levels are transient. If we just follow the “no” branch of the tree we reach the cluster C_5 . Its size indicates that the bulk of genes fall into this cluster. We believe that most of the genes that do not have a coordinated stress response fall into this cluster. Indicative of this is the cluster prototype, which shows no major changes in gene expression and has a large error.

4. Results

In this section, we present and evaluate the results of the analysis of time course gene expression data with PCTs. The expression data measures the response of yeast genes to different types of environmental stress and we first give a brief description of it. We then show how the produced PCT models can be interpreted in order to obtain biologically meaningful knowledge. We also discuss the similarity of the biological processes that are involved in the response to different types of stress. We finally present the results of experiments performed for assessing the predictive performance of the constructed PCTs.

4.1 Dataset description

For our experiments, we use the time-series expression data from the study conducted by Gasch *et al.*,⁷ which are publicly available. The purpose of the study is to explore the changes in expression levels of yeast (*Saccharomyces cerevisiae*) genes under diverse environmental stresses. The gene expression levels of around 5000 genes are measured at different time points using microarrays. The data is log-transformed and

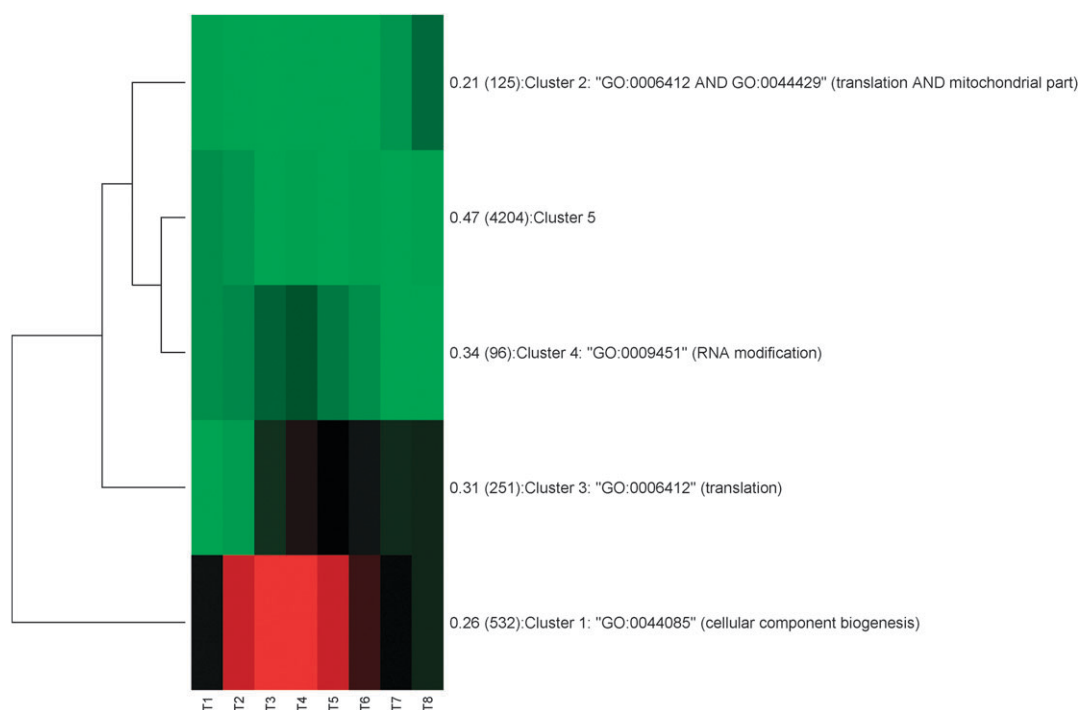


Fig. 4 Heatmap of the cluster prototypes and their accompanying descriptions from the PCT in Fig. 3. The first number on the right-hand side of the heatmap is the cluster's RMS error, the number in brackets is the cluster's size, the cluster's descriptions follow after the colon. "Cluster 2" contains genes that are involved in translation and whose protein products are a part of the mitochondria. These genes are significantly up-regulated. Cellular component biogenesis is strongly repressed, as evident on "Cluster 1". All of the clusters show a transient response to diamide, except "Cluster 5" which shows almost a constant temporal expression profile.

normalized based on the time-zero measurement of yeast cells under normal environmental conditions.

Various sudden changes in the environmental conditions are tested, ranging from heat shock to amino acid starvation for a prolonged period of time. We used a total of 10 datasets (different stress conditions) for our analysis. We perform a comparative analysis of the obtained descriptions from all of the datasets in section 4.4. For a more detailed discussion of the obtained descriptions (section 4.3) we considered four representative datasets, for different types of stressful conditions (temperature, chemical and starvation). Namely, we consider heat shock (from 25 to 37 °C), diamide treatment, DTT (dithiothreitol) exposure and nitrogen starvation.

From these original time series datasets, we construct extended datasets by including gene descriptions. We obtained the GO term annotations for each yeast gene from the Gene Ontology¹⁴ (version June, 2009). As the GO terms are structured in a hierarchy, we use both the *part_of* and *is_a* relations to include all relevant GO terms for each gene. To limit the number of features, we set a minimum frequency threshold: each included GO term must appear in the annotations for at least 50 of the 5000 genes.

4.2 Interpretation of PCTs for time course profiles

As explained in section 1, a PCT represents a hierarchical clustering of the time course data, where each leaf corresponds to one cluster. In Fig. 3, we present a sample PCT. For practical purposes, we show a small tree with just 5 leaves, obtained when yeast is exposed to diamide. We also show the

cluster centroids for each of the leaves. By following the path from the root of the tree to a leaf, we can obtain the description for each of the clusters.

For example, if we want to derive the description of cluster C_2 , we begin from the root GO term "GO:0044085", we follow the "no" branch, obtaining the description "GO:0044085 = no". We then add the "GO:0006412 = yes" and "GO:0044429 = yes" by following the "yes" branches ending up at cluster C_2 . So, the final description of cluster C_2 is the following conjunction: "GO:0044085 = no AND GO:0006412 = yes AND GO:0044429 = yes". This can be interpreted as follows: genes that are annotated by both "GO:0006412" and "GO:0044429", but not by "GO:0044085" are contained in cluster C_2 and have a temporal profile represented by the prototype of cluster C_2 .

It should be noted here that for our application only the positive branches of the tree are semantically meaningful. In a biological context, the description "GO:0044085 = no" is not very meaningful because it simply tells us that the genes in cluster C_2 are not annotated by that term. Therefore, to describe a cluster we only take the positive "yes" terms, which means that for describing C_2 we would only use "GO:0006412 = yes AND GO:0044429 = yes".

After deriving the descriptions from all of the clusters (except for cluster C_5), we represent them using a heatmap (Fig. 4). Each row in the heatmap represents a cluster prototype, the more intense the colours, the larger the up- or down-regulation of the genes contained in that cluster. Accompanying the rows, on the right-hand side, is the error of

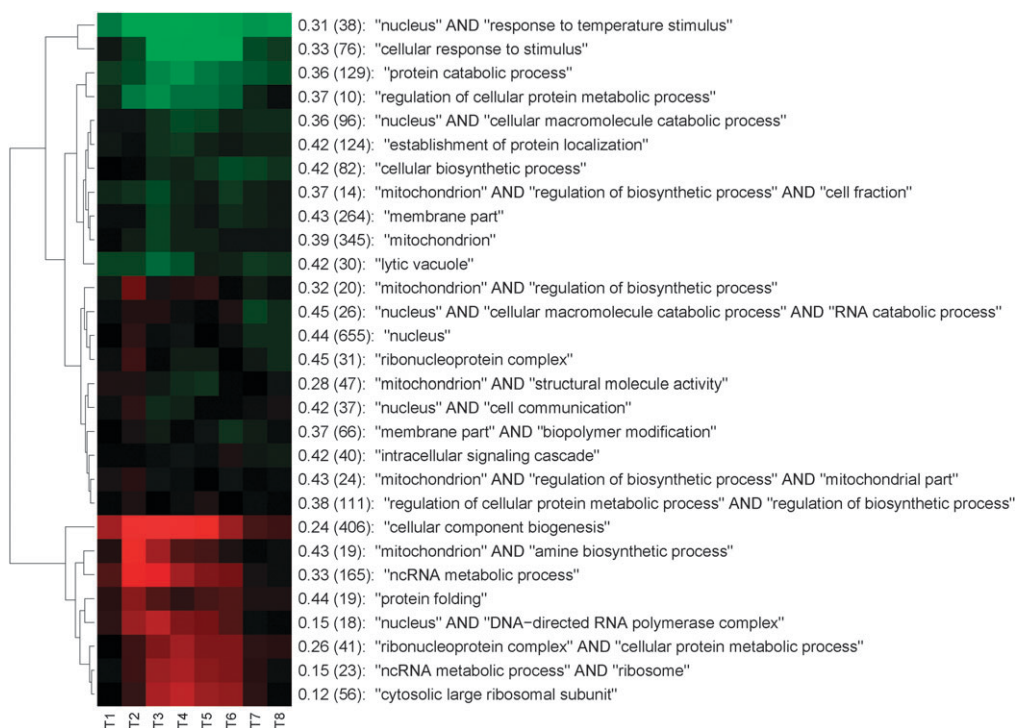


Fig. 5 When yeast is exposed to heat shock, several clusters of genes show significant, but transient changes in expression levels. According to the heatmap intensity, genes involved in response to temperature stimulus are most strongly induced. Down-regulated are genes involved in biosynthesis processes and genes that code for ribosomal proteins.

each cluster (RMSE, described in section 3.4), the cluster size and the cluster description. Note that the heatmap ordering of the cluster prototypes does not match the ordering produced by the

PCTs, but it is a permutation of it. This is for visualization purposes, in order to have all of the up- and all of the down-regulated cluster prototypes grouped together.

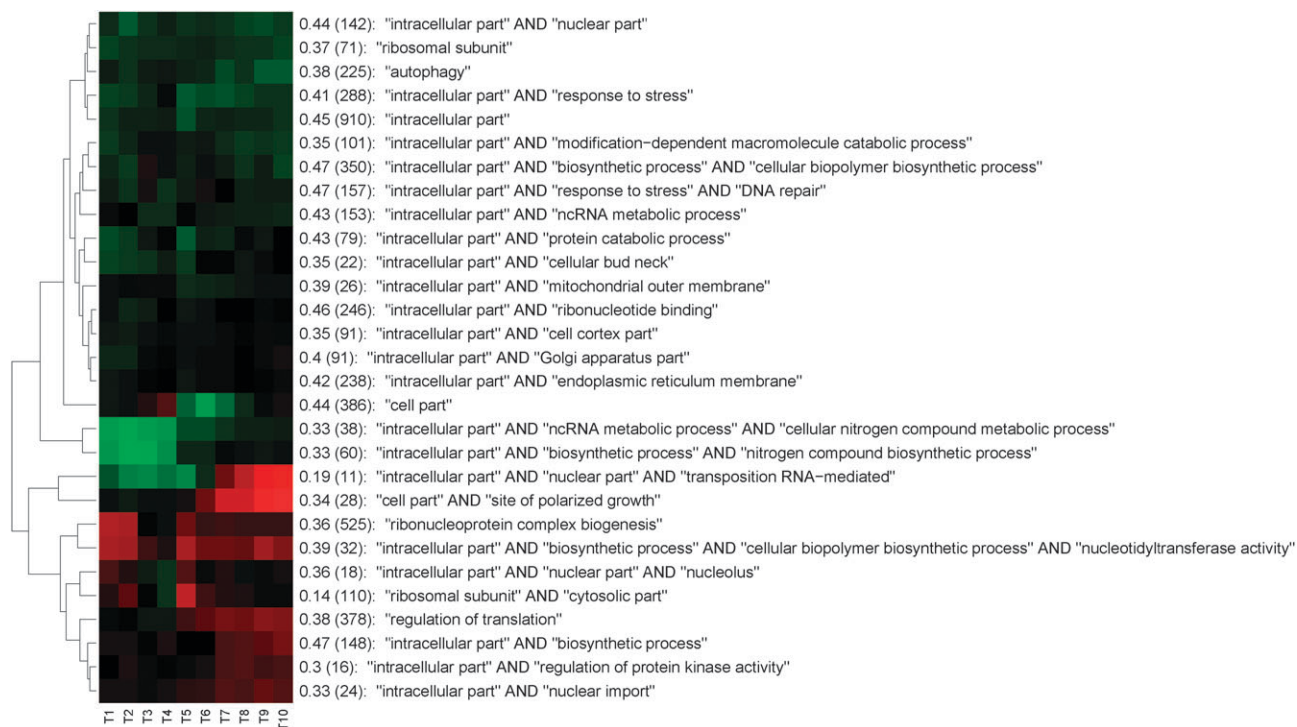


Fig. 6 Under nitrogen starvation conditions, there is more of a steady down-regulation of genes, rather than a transient pattern. Genes involved in nitrogen metabolism are slowly down-regulated as well as genes coding for ribosomal proteins.



Fig. 7 DTT treatment of yeast interferes with proper protein folding and changes the cellular redox state. Therefore, an up-regulation of genes involved in heat response and electron carrier activity is evident. More general biosynthetic processes and ribosomal proteins synthesis (nucleolus) are inhibited, *i.e.*, these genes are down-regulated.

4.3 Descriptions of yeast stress response clusters

We apply the procedure for deriving cluster descriptions from the previous section, on PCTs constructed for several datasets taken from a study of yeast stress response.⁷ While PCTs were

applied on all 10 datasets we only present in detail the results for four different stress conditions (heat shock, nitrogen starvation, diamide and DTT exposure). We present the final descriptions by using heatmaps given in Fig. 5–8.

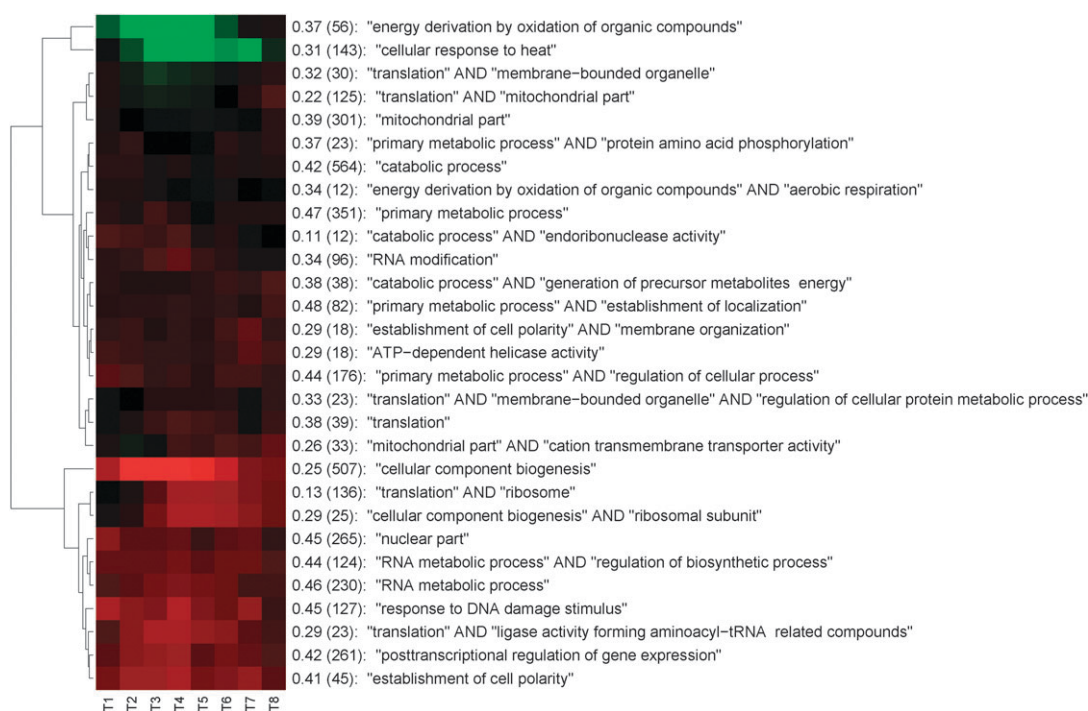


Fig. 8 Exposure to diamide causes a response similar to DTT treatment and heat shock, in terms of response to protein folding inhibition. Also oxidation of organic compounds is strongly up-regulated, while the down-regulation of biogenesis and ribosomal genes is also apparent.

We first consider the heatmap for the Heat Shock dataset, presented in Fig. 5. One can quickly identify two groups of temporal profiles in this figure: one that shows significant up-regulation and another one that shows significant down-regulation of genes. These significant changes are only transient in nature, meaning that genes first quickly react to the heat shock and then after an adaptation period go back to normal expression levels. This is an expected behavior, also noted in ref. 7, which shows that the predicted cluster prototypes are consistent with the biological reality.

From the induced genes, those involved in cellular response to stimulus, specifically temperature stimulus, show the most notable changes. In the repressed genes group, we can notice two slightly different groups with respect to the delay of response to heat shock. The first group that is quick to react, consists of genes involved in biogenesis and different biosynthetic and metabolic processes. A slight delay in down-regulation is exhibited by genes coding for ribosomal proteins, which is consistent with general stress response (ref. 7).

In contrast to heat shock, when yeast is subjected to nitrogen starvation, there is no transient temporal pattern present but more of a steady down-regulation of genes, as evident in Fig. 6. Genes involved in nitrogen metabolism slowly decrease their activity, while genes involved in cell growth are most significantly repressed. There is also a slight increase in the activity of autophagy genes.

The elicited response of yeast to DTT (dithiothreitol) is presented in Fig. 7. There is a small group of genes that is repressed over time. These are involved in general biosynthetic processes and genes that code for ribosomal proteins found in the nucleolus. Genes that were most induced are involved in electron carrier activity and genes that are a part of the general response of heat. This is the cell's response to the changed cellular redox state and to the inhibition of protein folding caused by DTT.⁷

Diamide exposure caused a response that can be seen as a combination between the response to heat shock and DTT. Genes involved in the heat shock response were induced (due to protein folding inhibition) as were genes involved in oxidation of organic compounds. As part of the general stress response, genes involved in cell component biogenesis were strongly repressed, as well as (with a small time-lag) genes coding for ribosomal proteins.

Note that we present descriptions derived from PCTs (in Fig. 5–8) from trees of size 60 (with 30 leaves/clusters). These usually consist of GO terms referring to general cell processes or locations. We chose this size as an optimal tree size, appropriate for viewing and with an acceptably low error (RMSE) (Fig. 10(a), (c), (e) and (g)). For obtaining clusters with more specific descriptions, one might consider larger trees.

4.4 Semantic similarity of biological processes involved in different types of stress

In the previous section, we presented the GO descriptions of the clusters of gene expression time profiles for four different stress conditions. We briefly discussed their similarities and differences in terms of the kind of cellular processes involved in stress response. Here, we focus on a more quantitative

analysis of the derived GO descriptions, where we also include a whole range of stress conditions.

To quantitatively compare the GO descriptions of the different clusterings, we use the semantic similarity measure between GO terms proposed by Wang *et al.*²⁷ Given two GO terms, this measure quantifies their functional similarity, by considering their common ancestors information from the Gene Ontology. By using the semantic similarity measure, we first determine the similarity between pairs of groups of GO terms, corresponding to pairs of descriptions of PCTs/clustering of yeast genes for different stressful conditions. We proceed by performing hierarchical clustering of the different stress types, in order to determine to which stressful conditions the yeast genes respond in the most functionally similar way (Fig. 9).

In Fig. 9, we can see that the cell response to heat shock is most similar to the response to DTT exposure and to the response when yeast is undergoing diauxic shift. This is due primarily to the response of genes to protein unfolding, which is the initial response to heat shock and DTT exposure. It is also due to the induction of genes involved in alternative carbon source utilization, which happens during diauxic shift and also as an aftermath of the heat shock.

Hyper and hypo osmotic shock are also grouped together, which is expected because they involve the response of the same set of biological processes, but they respond in an temporally inverse manner.⁷

Diamide is grouped with AA starvation and at a later stage with H₂O₂ exposure, which is expected due to its response being very similar to the response of these.⁷ Overall, we can

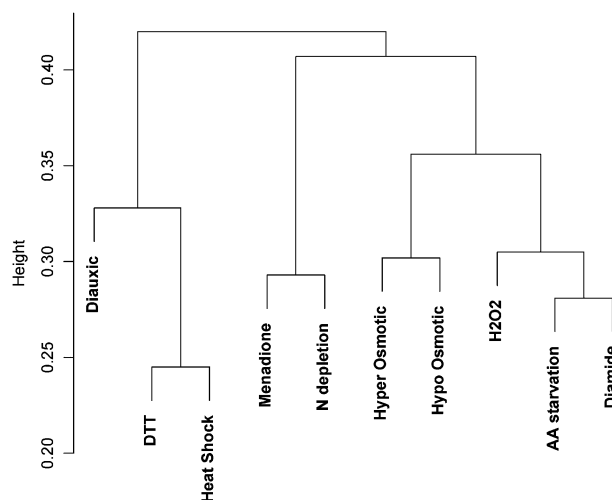


Fig. 9 In this figure, we present a dendrogram constructed according to the semantic similarity of the biological processes involved in response to different types of environmental stress. Heat shock is most similar to DTT exposure, which can be attributed to the protein unfolding which initially occurs in both types of stress. The other similarity to diauxic shift appears as a result of activation of processes for utilizing alternative carbon sources in the aftermath of heat shock. Hyper and hypo osmotic conditions are grouped together as they involve the same processes in response to the shock. Diamide is most similar to AA starvation and then to H₂O₂ exposure. Overall there is high similarity of all biological processes involved in different stress responses, which is indicative of the existence of a general stress response mechanism.

notice that the similarity (*i.e.*, distance) between the different stress conditions is relatively high (low), which implies that there is a commonality of cell responses to different types of stress, *i.e.*, a general stress response mechanism.⁷

4.5 Predicting time series with PCTs

We compare the PCTs built by Clus-TS (section 3.2) to a default predictor DEF that always predicts the overall training

set centroid. We estimate the RMSE of these predictions by using 10 fold cross-validation, as described in section 3.4. This means that when estimating the error for each fold, the training set contains approximately 4500 genes and the testing fold approximately 500 genes.

We first perform experiments for different maximum PCT sizes and we measure the respective RMSE of the corresponding PCTs. In Fig. 10((a), (c), (e) and (g)), we present the results for different values of the size upper bound. From the results,

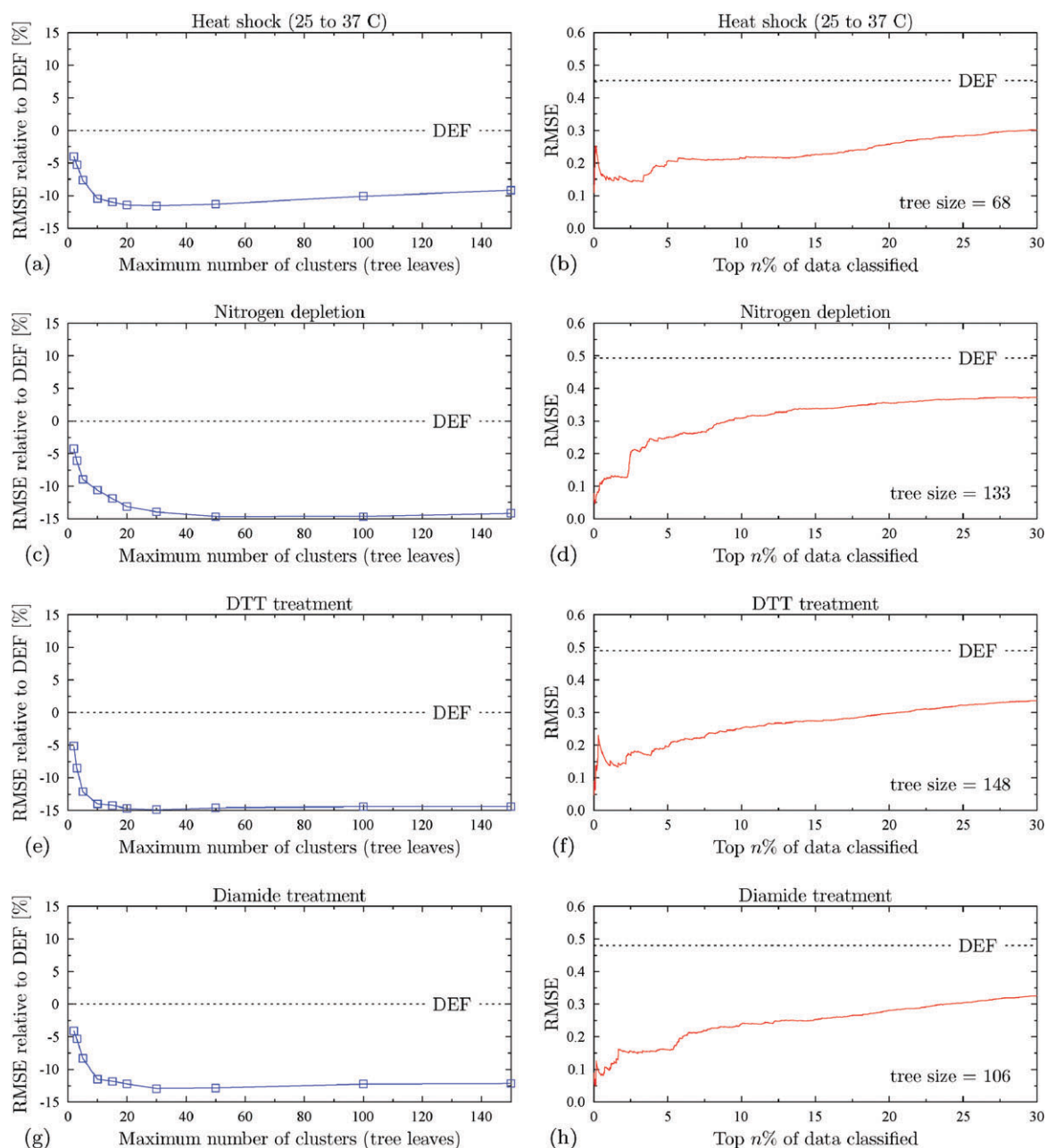


Fig. 10 A comparison of predictive error (RMSE) of PCTs for different number of clusters ((a), (c), (e) and (g)) and percentage of data classified ((b), (d), (f) and (h)). When increasing the maximum tree size (number of leaves) the RMSE decreases until the size of the tree reaches 20–30 leaves (*i.e.*, clusters). The maximal improvement in the overall RMSE, as compared to the default (DEF) error, is about 15%. This small decrease in the error is problem specific, *i.e.*, has a biological background: not all genes have a coordinated response to the different stresses. Therefore, the PCTs are only able to correctly predict the time-course profile of a limited number of genes. This is evident in (b), (d), (f) and (h). For about 5% of the genes, PCTs are able to correctly predict their time-course profile with a relatively low RMSE as compared to the default (DEF). DEF is the default predictor that always predicts the overall training set centroid.

we can see that the optimal tree size for the PCTs is around 30 leaves. But, as one can notice, the overall RMSE is still relatively high. We hypothesize that the overall high error (RMSE) is domain specific, *i.e.*, there is a biological explanation for it.

Namely, the PCTs cluster genes that are annotated by similar GO terms and have a similar response in expression level to a certain change in environmental conditions. One problem is that, as noted by Gasch *et al.*,⁷ only a subset of the genes (about 900) have a stereotypical response to environmental stress. That is, only a subset of the genes can be accurately clustered, whereas the other genes have an uncorrelated response. As a result, we hypothesize that the PCTs are able to accurately predict the time series of only a subset of the genes. We therefore perform the following experiment. Besides recording the predicted time series for each test set gene, we also record a confidence value for each prediction. We then sort the genes by confidence value and compute the RMSE of the top n percent most confident predictions. We use the training set RMSE of the leaf that made the prediction as a confidence estimate. This is similar to the approach used for generating a ROC curve for a decision tree.²⁸ We present the results in Fig. 10(b), (d), (f) and (h)). PCTs are obtained with the same parameters as before, except that we use validation set based pruning instead of specifying a size constraint on the PCTs. Clus-TS now uses 1000 genes of the original training set for pruning and the rest for the tree construction (as suggested by ref. 29). Simply selecting a PCT from Fig. 10(a), (c), (e) or (g) is unfair; it corresponds to optimizing the size parameter on the test set. The results show (Fig. 10) that more accurate predictions are obtained if we restrict the test set based on the confidence of the predictions. For example, if time course profiles are predicted for the 5% of genes with highest confidence then the RMSE decreases to about 50% of that of DEF. This is also shown in Fig. 3.

5. Conclusions

The typical approach to analyzing time-course expression data is to first group together genes with similar temporal profiles into clusters, which are then subsequently explained in terms of gene properties (such as GO annotations). We present a novel methodology for clustering time course profiles of gene expression data, which unifies the two steps of clustering and inferring a cluster description. The methodology produces a hierarchical clustering, called a predictive clustering tree, where each cluster is described by a conjunction of gene properties (such as GO terms).

There are several advantages of our approach over other analysis methods. First, we perform clustering and provide cluster explanations in a single step. The descriptions can use practically any gene-related information, although for our experiments we only included Gene Ontology terms. Second, in contrast to the usual distance measure used for clustering (typically correlation based), our approach uses a qualitative distance measure (QDM), which was specifically designed to deal with short time course data. This measure explicitly takes into account the temporal nature of the gene expression profiles, and captures mostly the similarity in the shape of

the time course data, which is very important for the application at hand. Third, the PCTs also enable the prediction of gene expression time profiles for genes based on their annotations (functions), which is usually not possible with other mainstream clustering approaches.

We apply the proposed methodology to cluster time course data representing yeast gene response to environmental stress. This is repeated for different types of stress producing different PCTs, thus producing different clusters and cluster explanations in terms of GO annotations. Upon close inspection, the explanations of the clusters were consistent with previously published biological results.⁷ Furthermore, clusters with similar descriptions under different stress conditions were identified, mainly related to biosynthesis and ribosomal proteins. The results demonstrate the usefulness of our method for analyzing time-course expression data.

Several directions for further work remain to be explored. We consider first and foremost extending our approach to a so-called multi-target approach. Instead of considering a single time course at a time, for different (stress) conditions, we can consider the responses to different kinds of environmental conditions simultaneously. The application of this would be, for example, discovering a common stress response pattern.³⁰ Instead of producing a separate PCT for each condition, we would obtain just one PCT model for all. Another direction of further research includes the identification of groups of genes with coordinated response. Namely, the hierarchical nature of the PCTs, besides producing compact clusters of “stress” response genes, also produces some clusters that contain genes without a coordinated response. To focus on clusters of genes with coordinated response, we plan to further investigate the use of the so-called predictive clustering rules⁵ for analyzing short time course data. Finally, we would like to apply the proposed approach to other time course gene expression data from different biological domains.

References

- 1 J. Ernst, G. J. Nau and Z. Bar-Joseph, Clustering short time series gene expression data, *Bioinformatics*, 2005, **21**(Suppl. 1), i159–i168.
- 2 H. Blockeel, L. De Raedt and J. Ramon, Top-down induction of clustering trees, in *15th Int'l Conf. on Machine Learning*, 1998, pp. 55–63.
- 3 L. Kaufman and P. J. Rousseeuw *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, 1990.
- 4 J. Struyf and S. Džeroski, Constraint based induction of multi-objective regression trees, *Lect. Notes Comput. Sci.*, 2006, **3933**, 222–233.
- 5 B. Ženko, S. Džeroski and J. Struyf, Learning predictive clustering rules, *Lect. Notes Comput. Sci.*, 2005, **3933**, 234–250.
- 6 T. W. Liao, Clustering of time series data—a survey, *Pattern Recognit.*, 2005, **38**, 1857–1874.
- 7 A. Gasch, P. Spellman, C. Kao, O. Carmel-Harel, M. Eisen, G. Storz, D. Botstein and P. Brown, Genomic expression program in the response of yeast cells to environmental changes, *Mol. Biol. Cell.*, 2000, **11**, 4241–4257.
- 8 Z. Bar-Joseph, G. Gerber, T. Jaakkola, D. Gifford and I. Simon, Continuous representations of time series gene expression data, *J. Comput. Biol.*, 2003, **10**, 341–356.
- 9 S. Datta and S. Datta, Comparisons and validation of statistical clustering techniques for microarray gene expression data, *Bioinformatics*, 2003, **19**(4), 459–466.
- 10 N. Heard, C. Holmes and D. Stephen, A quantitative study of gene regulation involved in the immune response of anopheline

- mosquitoes: An application of bayesian hierarchical clustering of curves, *J. Am. Stat. Soc.*, 2006, **101**, 18–29.
- 11 M. Ramoni, P. Sebastiani and I. S. Kohane, Cluster analysis of gene expression dynamics, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**(14), 9121–9126.
 - 12 A. Schliep, A. Schonhuth and C. Steinhoff, Using hidden Markov models to analyze gene expression time course data, *Bioinformatics*, 2003, **19**, 255i–1272.
 - 13 L. Wang, M. F. Ramoni and P. Sebastiani, Clustering short gene expression expression profiles, *Lect. Notes Comput. Sci.*, 2006, **3909**, 60–68.
 - 14 M. Ashburner, *et al.*, Gene Ontology: Tool for the unification of biology The Gene Ontology Consortium, *Nat. Genet.*, 2000, **25**(1), 25–29.
 - 15 M. Kanehisa and S. Goto, Kegg: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, 2000, **28**(1), 27–30.
 - 16 Z. Bar-Joseph, Analyzing time series gene expression data, *Bioinformatics*, 2004, **20**(16), 2493–2503.
 - 17 M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 14863–14868.
 - 18 J. Handl, J. Knowles and D. B. Kell, Computational cluster validation in post-genomic data analysis, *Bioinformatics*, 2005, **21**(15), 3201–3212.
 - 19 J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins and R. W. Davis, Significance analysis of time course microarray experiments, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 12837–12842.
 - 20 M. Mutarelli, L. Cicatiello, L. Ferraro, O. Grober, M. Ravo, A. Facchiano, C. Angelini and A. Weisz, Time-course analysis of genome-wide gene expression data from hormone-responsive human breast cancer cells, *BMC Bioinformatics*, 2008, **9**(Suppl 2), S12.
 - 21 J. Sese, Y. Kurokawa, M. Monden, K. Kato and S. Morishita, Constrained clusters of gene expression profiles with pathological-features, *Bioinformatics*, 2004, **20**, 3137–3145.
 - 22 J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann series in Machine Learning, Morgan Kaufmann, 1993.
 - 23 R. S. Michalski and R. E. Stepp, Learning from observation: Conceptual clustering, in *Machine Learning: an Artificial Intelligence Approach*, ed. R. S. Michalski, J. G. Carbonell, T. M. Mitchell, Tioga Publishing Company, 1983, vol. 1, pp. 331–363.
 - 24 L. Todorovski, B. Cestnik, M. Kline, N. Lavrač and S. Džeroski, Qualitative clustering of short time-series: A case study of firms reputation data, in *ECML/PKDD'02 Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning*, 2002, pp. 141–149.
 - 25 M. Garofalakis, D. Hyun, R. Rastogi and K. Shim, Building decision trees with constraints, *Data Min. Knowl. Discovery*, 2003, **7**(2), 187–214.
 - 26 H. Sakoe and S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, *IEEE Trans. Acoust., Speech., Signal Process.*, 1978, **26**(1), 43–49.
 - 27 J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu and C. Chen, A new method to measure the semantic similarity of go terms, *Bioinformatics*, 2007, **23**(10), 1274–1281.
 - 28 C. Ferri, P. A. Flach and J. Hernández-Orallo, Learning decision trees using the area under the ROC curve, in *19th Int'l Conf. on Machine Learning*, 2002, pp. 139–146.
 - 29 L. Torgo, A comparative study of reliable error estimators for pruning regression trees, *Lect. Notes Comput. Sci.*, 1998, **1484**.
 - 30 L. Wang, M. Montano, M. Rarick and P. Sebastiani, Conditional clustering of temporal expression profiles, *BMC Bioinformatics*, 2008, **9**, 147.