Exploratory statistical analysis of multi-species time course gene expression data

Eng, Kevin H.
University of Wisconsin, Department of Statistics
1300 University Avenue, Madison, WI 53706, USA.
E-mail: eng@stat.wisc.edu.

Kvitek, Dan
University of Wisconsin, Department of Genetics
425 Henry Mall, Madison, WI 53706, USA.

Wahba, Grace
University of Wisconsin, Department of Statistics
1300 University Avenue, Madison, WI 53706, USA.

Gasch, Audrey
University of Wisconsin, Department of Genetics, Genome Center of Wisconsin
425 Henry Mall, Madison, WI 53705, USA.

Keleş, Sündüz[1]
University of Wisconsin, Departments of Statistics and of Biostatistics and Medical Informatics
1300 University Avenue, Madison, WI 53706, USA.
E-mail: keles@stat.wisc.edu.
[1] *Corresponding author.*

Multi-species time course gene expression data

The major strengths of the gene expression platforms lie in their ability to illuminate high dimension, highly correlated, time varying processes in the cell. While the conventional gene expression analysis can associate individual genes with conditions of interest, time course gene expression analysis ties together genes in functional groups and associates them with biological processes. Such processes can be natural cycles as in the well known Spellman yeast cycle data (Spellman *et al.*, 1998), or they can be responses to stimuli as in the Gasch yeast environmental stress data (Gasch *et al.*, 2000). In these analyses, the assumption is that genes which vary over time and which are correlated with one another represent functional groups and the goal is to uncover these biological clusters.

The multi-species extension of such methods allows the discovery and investigation of functional groups in different species. It is of interest to determine whether the characterized responses of individual genes and correlated groupings of genes in the reference species have a convergent or divergent response in other species. Ideally, this analysis might uncover differential, species specific time patterns which could be tied to other genetic or biochemical regulatory information (Rifkin *et al.*, 2003). One might consider going as far as overlaying a phylogenetic map to make evolutionary inferences (Gu, 2004).

Dedicated multi-species time course experiments are relatively novel and the results of an analysis

are hard to interpret (Gilad *et al.*, 2006). Therefore, the goal of this paper is to present an exploratory analysis pipeline for the multi-species experiment, outlining the main difficulties in extending the single species time course analysis. We pay particular attention to presenting the results so as to best inform downstream biological analyses.

Motivating data example

Our motivating dataset is a multi-species extension of the Gasch study (Gasch *et al.*, 2003), which identifies a number of genes in the yeast *Saccharomyces cerevisiae* that react to varying stresses including heat shock. These environmental stress response genes (ESR) have a transient response: activating and eventually returning to a baseline level of expression. The multi-species study measures the response to heat shock and investigates the conformity of response in different strains of yeast.

There are six points in the time course (plus a baseline measurement) and two replicate time courses for each of four strains (48 arrays total). All experiments were carried out using the same type of spotted cDNA array which measures 5407 genes. The recorded data are base 2 log ratios of each time point over the time course baseline. Thus the image plots in the analysis appear green (negative) when a gene is repressed versus the baseline measurement and red (positive) when a gene is induced versus the baseline measurement. Dark parts of the plots correspond to small differences from the baseline measurement.

There are two technical points which deserve mention but we do not discuss further in this paper. First, for increasingly distant species (evolutionary-wise), gene orthologs are increasingly rare. While the data presented here are very closely related (they are all strains of S. cerevisiae) and thus avoid the problem, this will not be generally true of multi-species data. Second, the typical downstream analysis for a gene expression clustering analysis consists of a validation step where discovered clusters are matched to gene ontology (GO) functionally annotated categories. In this multi-species setting it is not clear that a functional group of genes defined in one species have a consistent interpretation in another. It may be a useful analysis, in fact, to pick specific GO clusters and determine how the analysis portrays these groups.

Methods

ANOVA model and parameterization. Since our goal is to model the mean gene expression profiles of multiple species across time, we fit an ANOVA model to each gene on time and species factors. Consider the effects model for a single gene (subscript suppressed):

$$Y_{tsk} = \mu + \tau_t + \beta_s + \gamma_{ts} + \epsilon_{tsk}$$

for time $t = 1, \ldots, T$, species $s = 1, \ldots, S$, and replicate $k = 1, \ldots, K$. In the single species setting, we wish to group genes with similar time dependent responses into functionally related clusters. We consider specific genes which differ by a constant amount but which show the same time pattern, relative to their baseline measurements, to display a similar response over time and believe that these should be grouped together. In the multi-species setting, with an obvious reference species, the parameters of an effects model with set-to-zero constraints have biologically useful interpretations:

1. The grand mean, μ, is the baseline expression (at the first time point) for the reference species while the species coefficients, β, are differences in baseline expression for the other species.

2. μ together with the time coefficients, τ, represent the gene's time profile in the reference species, while the γ coefficients capture the species specific change in profile.

3. The τ's themselves are deviations from baseline for the reference species; likewise the γ's are deviations for other species.

Subgroupings. Having fit a model to each gene, we consider the Type I F-tests. First, genes which show no significant time main effect (i.e., in the τ's) are dropped from the analysis. We apply Benjamini and Hochberg's (Benjamini and Hochberg, 1995) linear step up procedure to control the false discovery rate (FDR) for the time effects only. We use the remaining tests to partition the remaining pool of genes as in the following table.

	Category	F-tests
	No Time Dependence	$\tau = 0$
Subgroups	No Species Effect	$\tau \neq 0, \beta = 0, \gamma = 0$
	Species Effect	$\tau \neq 0, \beta \neq 0, \gamma = 0$
	Entangled Effect	$\tau \neq 0, \beta \neq 0, \gamma \neq 0$

The "No Species Effect" subgroup contains the genes for which the time pattern is the same across all species, or for which there is only a time effect. Therefore, the relevant coefficients for the No Species group are μ and the τ's. In the "Species Effect" subgroup the pattern varies only by a fixed constant across species. The "Entangled Effect" subgroup represent genes with interaction effects, for which the time pattern differs from the reference species in at least one other species. Also, some genes may show significant interaction effects but no species effect. In this data set, the number is small enough ($\leq 1\%$ or about 50 genes) to investigate individually.

Recalling that the goal of the single species analysis was to find functional groupings, that is, genes with similar time patterns, it is somewhat useful to think of the No Species subgroup as a subset of the usual single species setting. Accordingly, one expects that the standard hierarchical clustering methods can be applied with the usual functional groupings interpretation. Likewise, conditional on the species effects, the time patterns of the Species subgroup could be similarly grouped as if they were a single species. The interpretation of such clusters would not be exactly the same as in the single species analysis, and we offer an alternative analysis and pay some attention to the interpretation of the results in the following sections.

Sign Patterns. We have it in mind to group genes using their coefficient information, in order to make good use of the ANOVA parameterization. Within each subgroup, the variety of signals is difficult for standard clustering methods (pam, mclust, k-means, (Duda and Hart, 1973)) to handle. Visually, we want to see similar groupings of red and green patterns on the image plots, and this intuition corresponds to considering the signs of the coefficients.

For each subgroup, we tally the frequency of the unique patterns of the coefficients and select only the patterns present in at least 10 genes. We interpret each subpattern as a cluster and describe each cluster with a multivariate normal density fit on the coefficient estimates. For each gene, we evaluate a measure of cluster association, a Membership Assignment Function (MAF), based on the posterior probability of cluster membership. For clustering variable $Z_g^{(l)}$, gene g and subgroup l with

coefficients $X_g^{(l)}$, and K_l detected patterns,

$$Pr(Z_g^{(l)} = z \mid X_g^{(l)}) = \frac{Pr(X_g^{(l)} \mid Z_g^{(l)} = z)Pr(Z_g^{(l)} = z)}{\sum_{z'} Pr(X_g^{(l)} \mid Z_g^{(l)} = z')Pr(Z_g^{(l)} = z')}, \quad z' = 1, 2, \ldots, K_l$$

Genes whose patterns are infrequently observed are assigned to the clusters according to the MAF. It is worth noting that some genes may be difficult to assign, and it is natural to assign genes with a maximum MAF value of at most 0.50 to a "Null Cluster." Such genes are not uninteresting, they are merely hard to associate with other genes. These singletons, especially if they have large average expression, ought to be investigated individually.

Discovering patterns using the signs, while informative, can result in coarsely defined clusters. Genes with the same sign profile but significantly different mean values may not appear in separate clusters and some consideration may be given to how to further divide clusters. Using the signs of the coefficients also aids in the interpretation of the resulting clusters; while we have continuous estimates for the cluster means, we can fall back on the interpretation of the original sign pattern.

No Species. The resulting clusters group together genes which have similar time profiles. The following example patterns are generally induced/red ($\text{sign}(\hat{\mu}) = 1$) with different time patterns.

Sign of:	μ	τ_2	τ_3	τ_4	τ_5	τ_6
Cluster 1	1	1	-1	-1	-1	-1
Cluster 2	1	-1	-1	-1	-1	-1
Cluster 3	1	1	1	-1	-1	-1

Cluster 1 contains genes which increase in intensity or peak near time point 2 (T2) before dropping to a level below baseline, while cluster 2's genes are less induced immediately following the baseline measurement. Cluster 3 peaks around time point 3 before reducing in intensity.

Species. We consider the coefficients in two levels. First, using the species coefficients (the β's) groups together genes which have similar baseline patterns across the species. For example, for strains K9, M22 and RM11a,

Sign of:	β_{K9}	β_{M22}	β_{RM11a}
Species Level Cluster 1	1	1	-1
Species Level Cluster 2	1	-1	-1

Cluster one contains species effect genes which are, at the first time point, more induced in K9 and M22 versus the reference species and more repressed in RM11a versus the reference species. Likewise cluster two contains genes induced in K9 and repressed in M22 and RM11a.

Within each of these groupings, we interpret the μ and τ estimates as in the No Species grouping. These clusters find prominent time patterns conditional on a specific baseline pattern. Some time patterns, therefore, are represented in multiple clusters. For example, one from each of the

Sign of:	β_{K9}	β_{M22}	β_{RM11a}	μ	τ_2	τ_3	τ_4	τ_5	τ_6
Cluster 3	1	1	-1	1	-1	-1	-1	-1	-1
Cluster 6	1	-1	-1	1	-1	-1	-1	-1	-1

clusters above:

Both clusters have the same time profile, and differ only in species specific behavior. In effect, this approach adds the species information to the reference species time information by repeating the single-species analysis . Note that despite the two-level interpretation, the MAF is evaluated jointly, on all the relevant coefficients which allows for the possibility that genes with a specific sign pattern may actually be closer to a different cluster when we account for the estimated coefficient values.

Entangled. In this data set, the interpretation for entangled genes is unsettled. Some patterns exist, but there are too few genes (228) and too many patterns (79) for obvious clusters to emerge. Four well-represented patterns appear but the coarseness of the sign patterns results in clusters which are too difficult to summarize. For this analysis, we associate genes in the entangled group with the nearest species cluster using the membership assignment function ignoring the interaction and without refitting the model.

Results

Controlling the FDR at level 0.001, 2495 of the 5407 genes have significant time patterns. The following table relates the breakdown into subgroups with the remaining tests at an uncorrected level of 0.001. Recall that there are six time points and four species. We find 49 well-represented patterns, ignoring the four entangled patterns, which we can interpret guided by the signs of the coefficients as described previously.

	No Species	Species	Entangled
no. genes	1021	1218	228
no. coefficients	6	9	24
no. unique patterns	26	328	79
no. patterns, $\geq$10 genes	10	39	4

Our primary tool for summarizing the clustering results is the image plot. On these plots, rows correspond to individual genes and groupings of rows separated by white space are clusters of genes. Within the clusters, genes are ordered by MAF, largest at the top. Arrays are laid out in the columns where groups of columns are replicate time courses. In each time course, the left-most column is time point one and the right-most is time point six. All plots presented here show the raw data values for which positive log ratios over baseline are red and negative log ratios are green. Any ratio with a absolute fold change of less than 2 (or 1 on the $\log_2$ scale) is colored black for contrast.

Example: Discovered Clusters. Figure 1 is an image plot which shows a previously undiscoverable species specific pattern. The topmost cluster shows the pattern as discovered in the No

Species group. With the addition of species information in the remaining clusters, an exclusively strong M22 effect becomes prominent. The entangled subgroup is not represented.

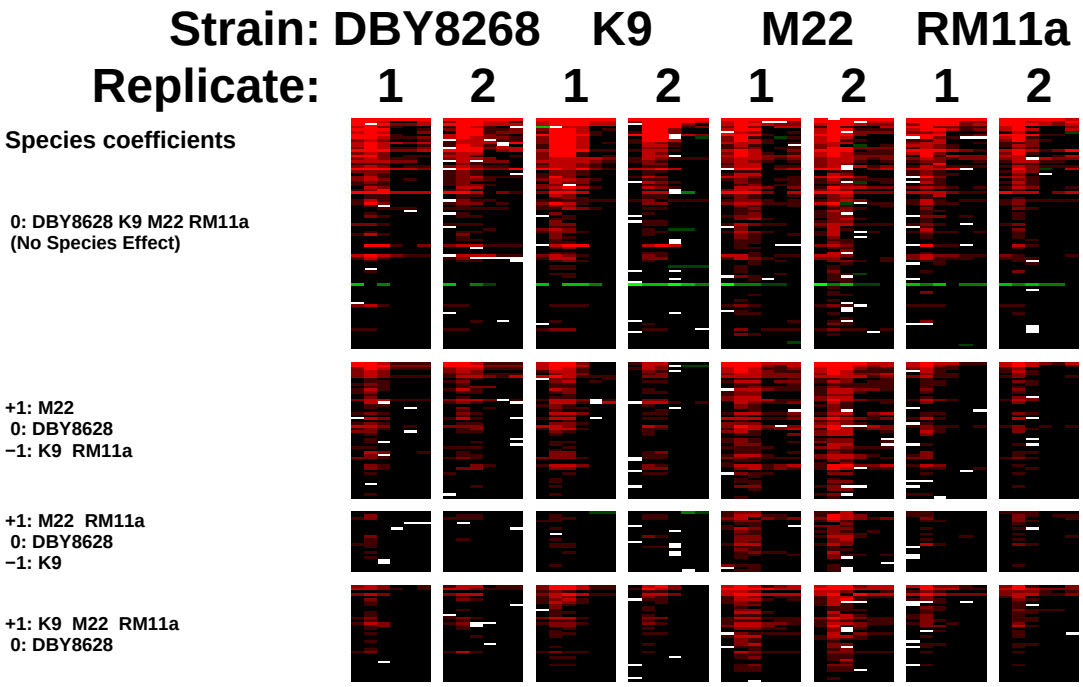

Figure 1: Clusters with time sign pattern (1 1 1 -1 -1 -1). This set of clusters highlights the advantages of species level information, compare the M22 pattern in the No Species Effect cluster versus the others.

Example: ESR genes. We were particularly interested in following the behavior of a set of ESR genes known to react to heat shock stress (Gasch *et al.*, 2003). The ESR genes comprise three functional groups defined below and cross-tabulated with their subgroupings in this analysis.

ESR Genes

	iESR	PAC	RP
No Time Effect	76	34	0
No Species	83	111	70
Species	91	85	41
Entangled	11	10	7

The subgroupings are very informative even without further clustering. 83 of the known induced ESR (iESR) genes have a completely consistent pattern across the other species while 91 experience a shift in baseline. For the 11 entangled genes, the functional response may be different. In Figure 2, we have chosen a single time pattern within the iESR genes to illustrate how interpreting the clustering results as disparate functional groups may be inappropriate. All of these iESR genes share a common time pattern and differ in their species effects. The top cluster contains iESR genes with the same

pattern across all species while the third cluster shows a clearly brighter (more red) pattern in the M22 strain. The last cluster contains genes with roughly the same brightness in K9 and the reference strain DBY8268 but where M22 and RM11a have roughly the same intensity. This example illustrates the proof-of-principle that our method can identify previously unrecognized subgroups of the ESR genes based on strain-specific responses, pointing to differences in their function, regulation, or evolution.

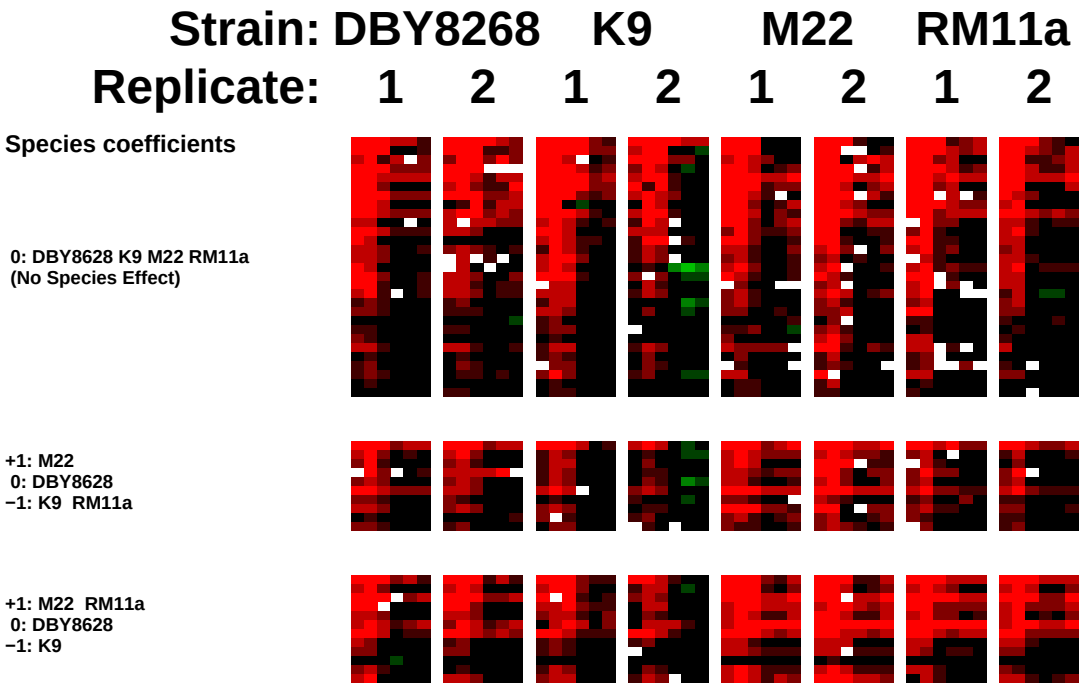

Figure 2: iESR genes with time pattern (1 1 -1 -1 -1 -1). New behavior for previously characterized genes includes genes with uniformly stronger responses in the non-baseline species.

Discussion

We have presented an exploratory analysis pipeline for multi-species time course gene expression data. We justify the structural assumptions in our procedure by first ruling out the results of direct applications of hierarchical clustering as too difficult to interpret. In other words, such an approach results in a clustering structure which requires extensive human evaluation to tease out the type of effects that we discuss in this paper. The use of subgroups and sign patterns make for clear, if coarse, cluster interpretations. Since we utilize cluster-specific multivariate normal densities to summarize the clusters and perform the cluster assignments, one may suggest applying a model-based clustering, e.g., mclust (Fraley and Raftery, 2002) on the raw coefficients. Our initial attempts indicate that a vanilla application of `mclust` produces uninformative clusters within subgroups. However, a more structured clustering model which carefully parameterizes the effects of interest might lead to well defined and biologically interesting clusters (e.g., Jornsten and Keleş, 2007).

There are several other important aspects of multi-species time course gene expression data that we have not addressed here. First, our analysis is constrained within a linear fixed effects model. However, given the gene and species level heterogeneities in the response, a linear mixed effects model (Luan and Li, 2003) might be more appropriate. Furthermore, we have treated the identification of conservation subgroups (No Species Effect, Species Effect, and Entangled Effect) separate from the clustering process. A model that is capable of simultaneously capturing both the subgroups and the clustering structure might be more powerful. Our analysis indicates that, for the ease of downstream interpretation, it is important to carefully parameterize the models and potentially utilize parameter selection during the process of clustering. Such an approach would potentially generate readily interpretable clusters. Finally, methods that can rigorously exploit the phytogenetic information of the species under study might provide further power in detecting convergent and divergent patterns of gene expression across multiple species.

REFERENCES (RÉFERENCES)

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. (1998) "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization". Molecular Biology of Cell 9(12):3273-97.

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. (2000) "Genomic expression programs in the response of yeast cells to environmental changes". Molecular Biology of Cell 11(12):4241-57.

Rifkin SA, Kim J, and White KP. (2003) "Evolution of Gene Expression in the Drosophila melanogaster subgroup". Nature Genetics 33: 138-144.

Gu, X. (2004) "Statistical framework for phylogenetic analysis of expression profiles". Genetics, 167:531-542.

Gilad Y, Oshlack A, Smyth GK, Speed TP, and White KP. (2006) "Expression profiling in primates reveals a rapid evolution of human transcription factors". Nature 440(9): 242-245.

Benjamini Y, Hochberg Y. (1995) "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". Journal of the Royal Statistical Society B 57(1):289-300.

Duda RO and Hart PW. (1973) "Pattern classification and scene analysis". John Wiley & Sons.

Fraley C and Raftery AE. (2002) "Model-based clustering, discriminant analysis, and density estimation". Journal of the American Statistical Association 97:611-631 (2002).

Jornsten JR and Keleş S. (2007) "MIXL, Multi-level mixture modeling" Under revision for Biostatistics.

Luan Y and Li H. (2003) "Clustering of time-course gene expression data using a mixed-effects model with B-splines". Bioinformatics 19 (4): 474-482.