Published in final edited form as:

Stat Med. 2013 October 15; 32(23): 4057–4070. doi:10.1002/sim.5850.

Modeling Considerations for Using Expression Data from Multiple Species

Elizabeth Siewert^{1,2,3} and Katerina J. Kechris¹

Katerina J. Kechris: katerina.kechris@ucdenver.edu

¹Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver

Abstract

Although genome-wide expression data sets from multiple species are now more commonly generated, there have been few studies on how to best integrate this type of correlated data into models. Starting with a single-species, linear regression model that predicts transcription factor binding sites as a case study, we investigated how best to take into account the correlated expression data when extending this model to multiple species. Using a multivariate regression model, the phylogenetic relationships among the species were accounted for in two ways: 1) a repeated-measures model, where the error term is constrained, and 2) a Bayesian hierarchical model, where the prior distributions of the regression coefficients are constrained. We show that both multiple-species models improve predictive performance over the single-species model. When compared with each other, the repeated-measures model out-performed the Bayesian model. We suggest a possible explanation for the better performance of the model with the constrained error term.

1 Introduction

Microarrays measure the expression of thousands of genes simultaneously and were first introduced in the mid 1990's [1]. Initially, only expression data were collected to study, for example, the differential expression patterns of genes in case-control experiments. Later, gene expression data were combined with other types of data to address other interesting biological problems (e.g., the phenotypic diversity of gene expression as it relates to genome sequences).

More recently, expression data taken across multiple species have become more common and have been used to study a wide range of biological processes. These multi-species data have been used to study the evolution of gene expression [2–8] and the evolution and conservation of gene networks [9–11] across species. Microarray expression data across multiple species have also been used to predict gene function [12–15]. Stuart, *et al.*, for example, compared gene expression from the widely diverged organisms of humans, flies, worms, and yeast across multiple experimental conditions, and reported the prediction and

³Corresponding author siewertb@gmail.com.

²Currently with Statistically Speaking Consulting

experimental confirmation of genes previously unknown to be involved in important biological processes, such as cell cycle, secretion, and cell signaling. A recent review of cross-species analyses of microarray expression data is Lu, *et al.* [16].

Recently, it was shown that drawing information from multiple species improves the predictive accuracy of TFBSs [17–20]. For example, Kechris and Li [18] integrated cross-species sequence data into the single-species linear regression model proposed by Bussemaker *et al.* [21] in their pioneering method called REDUCE for Regulatory Element Detection Using Correlation with Expression. They showed that this improved TFBS prediction compared to only using sequence data from a single species. Expanding this concept further, Siewert and Kechris [22], developed a method that also included cross-species expression data along with cross-species sequence data and incorporated a covariance matrix imposed on the error term to account for the phylogenetic relationships among the species. This model, called multiple-species REDUCE (m-REDUCE) demonstrated an improvement of TFBS prediction over using expression and sequence data from just a single species.

With the introduction of expression data from multiple species, the question arises how to best account for the correlated nature of these data in statistical models. In this work, using the TFBS prediction model as a case study, we focus on where the correlation structure should be imposed in a multivariate, multiple-linear model. We evaluated the predictive performance of the multivariate model based on two ways of constraining it. We used the previously published m-REDUCE model which imposed the phylogenetic constraint on the variance of the error term and a model described here, b-REDUCE, which imposed the constraint on the variance among the regression coefficients. We found that while both models improved performance over the single-species algorithm, m-REDUCE tended to perform better than b-REDUCE. We present evidence for why it is better to impose the phylogenetic constraint on the error term rather than on the regression coefficients.

2 Methods

2.1 Data Sets

2.1.1 Expression Data—As part of their study of the evolution of gene expression, Tirosch *et al.* [8] generated time courses of microarray data in response to a number of stresses across four species of yeast, *Saccharomyces cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, and *Saccharomyces kudriavzevii*. We used the 10 minute time point from the heat-shock data set. Each data point was the ratio of the expression at 10 minutes relative to the expression at time 0 (when log-phase growing cells were subjected to a sudden increase in temperature from 30°C to 37°C) for every gene. The data were log₂ transformed, corrected for technical biases, and normalized by the authors [8]. We obtained these data from the NCBI GEO website (<http://www.ncbi.nlm.nih.gov/geo/> Data Sets Accession Code: GSE3406).

2.1.2 Sequence Data—The upstream promoter sequences for the genes in the four species were taken from the work of Cliften *et al.* [23] and Kellis *et al.* [20]. Although the model used unaligned sequences, the Information Content evaluation used aligned sequences

(see Section 2.3). These sequences were aligned using ClustalW [24] and the average alignment length was ~800 base pairs. We used the 3347 genes that had sequence and expression information for all four species.

2.1.3 Validation Data Sets—We used two independent data sets for evaluation purposes. The first, a microarray expression time course data set, was a study of expression patterns due to environmental changes in *S. cerevisiae* [25]. We used the heat-shock response data at the 15 minute time point after a temperature shift of 25°C to 37°C. We obtained these data from http://genome-www.stanford.edu/yeast_stress/data.shtml at the *Complete Dataset* link.

The second data set was the result of a Chromatin ImmunoPrecipitation - Microarray chip (ChIP-chip) assay. Harbison *et al.* [26] measured the binding strength of 203 TFs to DNA under various conditions in *S. cerevisiae*. We used these binding strength measures under all available conditions of the seven TFs that the authors stated were associated with the heat-shock response (18 data sets). The number of genes that had both sequence and expression data that also had ChIP data ranged from 3015–3347. These data files were downloaded from the Young Lab Regulatory Code website (http://jura.wi.mit.edu/young_public/regulatory_code) at the *Download Location Data* link. The *Excel* link under the *Data used for analysis (P values)* header can be used to extract the data. Once the data are extracted, use the *pvalbygene forpaper abbr.xls* file.

2.2 Models

2.2.1 Multiple Species Models—The binding of a transcription factor (TF) to its respective DNA recognition sequence, called a TFBS, is a key first step in gene expression. Based on this, Bussemaker *et al.* [21] assumed that every occurrence of a motif that functions as a TFBS will have an additive effect on the expression of any gene with that motif in its promoter region. They also assumed that each motif will contribute its own unique effect to the gene expression, but for a given motif, each occurrence of that motif will contribute the same additive affect across all genes.

Their method, called REDUCE, utilizes the sequence and expression information from all genes in a single species. It searches all possible motifs of length 5-7 nucleotides in the promoter region of every gene and chooses the motif whose number of occurrences is most significantly correlated with gene expression. That motif is entered into the model, the residual expressions for all genes are calculated, and the next motif whose count is most significantly correlated with the residual expressions is then found and incorporated into the model. This continues in an iterative manner until there are no more significant motifs found. Mathematically, their model was defined as ¹

¹The original paper does not explicitly define an error term, therefore we express their model in terms of $E[Y_g]$.

$$E[Y_g] = \beta_0 + \sum_{m \in M} \beta_{1m} X_{mg}$$

Y_g is the expression for gene g

X_{mg} is the number of occurrences of the motif m in the promoter region of gene g

M is the set of significant motifs

β_0 is the baseline expression value for all genes and

β_{1m} is the regression coefficient for motif m

β_0 and β_{1m} are the same for all genes.

Our multiple-species models are based on an extension of this single-species, simple linear regression model – replacing a single response variable with a vector of response variables, and a single count with a matrix of counts. Specifically,

$$\vec{Y}_g = \vec{\alpha} + \mathbf{X}_g \vec{\beta} + \vec{e}_g$$

$\vec{Y}_g$ is an $s \times 1$ vector of expression values of the s species for gene g

$\vec{\alpha}$ is an $s \times 1$ vector of intercepts

$\mathbf{X}_g$ is an $s \times k$ design matrix of the count of occurrences of the k explanatory motifs for the s species in the promoter region of gene g

$\vec{\beta}$ is a $k \times 1$ vector of regression coefficients for motif₁, motif₂, ..., motif_k ($\beta_1, \beta_2, \dots, \beta_k$)

$\vec{e}_g$ is an $s \times 1$ vector of errors for gene g .

In both models, we count the occurrences of every possible motif in the *unaligned* promoter regions of all genes in all species and, once the most significant motif is found, we include that motif as a covariate, and iteratively search for the next most significant motif in a forward selection manner.

So that similar sequences due to closely related species are not mistaken for functionally conserved sequences across species, we impose a correlation matrix that reflects the phylogenetic relationships among the species. There are two places where this phylogenetic matrix can be imposed – on the error term ($\vec{e}$) or on the regression coefficient terms

($\vec{\alpha}$ and $\vec{\beta}$).

Phylogenetic Correlation Matrix: The expression data set from Tirosh *et al.* [8] was generated from four species of the *sensu stricto* yeast group. To impose a correlation matrix that is independent of this data, we obtained the topology of the phylogenetic trees among five members of the *sensu stricto* group from [27–29]. The branch lengths of the phylogeny have been published [20], but did not include *S. kudriavzevii*. Therefore, we used an average branch length between *S. mikatae* and *S. bayanus* for the branch length of *S. kudriavzevii*. Figure 1 shows the published phylogenetic tree together with the branch lengths which is

used to determine the distances between species pairs. The phylogenetic correlation matrix, $\mathbf{D}$, was calculated as $1 - \text{phylogenetic distance matrix}$ [30].

Phylogenetic Constraint on Error Term (m-REDUCE): This model (m-REDUCE), a multivariate, multiple linear regression model, is described in detail in Siewert and Kechris [22]. Briefly, m-REDUCE is a repeated measures model where we treated the four expression measurements for each gene (one from each species) as four repeated measures of the expression of that gene across evolutionary time. Therefore, the phylogenetic correlation matrix was imposed on the error term, but the model also allowed for additional within species variance. Specifically

$$\vec{e} \sim \mathcal{N}(\vec{0}, \mathbf{R})$$

where

$\mathbf{R}$ is the covariance matrix of the error term
 $\mathbf{R} = \lambda_1 \mathbf{D} + \lambda_2 \mathbf{I}$
 $\mathbf{D}$ is the phylogenetic correlation matrix
 $\mathbf{I}$ is the identity matrix
 and λ_1, λ_2 are free parameters.

We assumed that each gene is activated or repressed in a similar manner for all four species, so we required that the regression coefficients be the same for all genes *and* all species (i.e., all 4 entries in $\vec{\alpha}$ are the same and there is only one β_i per motif i in $\vec{\beta}$). We chose the most significant motif at each iteration to be put into the model in a forward selection manner by testing the slope of the regression line for every possible motif using a t-distribution to compute the p-value for the two-sided test ($H_0: \beta_i = 0$ vs. $H_1: \beta_i \neq 0$). Significance was measured at the 0.01 level after a Bonferroni correction for the number of motifs tested. This was implemented using SAS 9.1.3 proc mixed, and specifying the correlation matrix, $\mathbf{R}$, in the repeated statement.

Phylogenetic Constraint on Regression Coefficients (b-REDUCE): The second model (b-REDUCE) uses a Bayesian hierarchical approach and is described in detail in the Ph.D. thesis of Siewert, 2010². We relax the requirement that the regression coefficients be the same for all genes *and* all species, and only require that it be the same for all genes for each predicted TFBS, but not for all species (i.e., the 4 entries in $\vec{\alpha}$ and $\vec{\beta}$ can vary according to the phylogenetic matrix). That is, rather than treating the four species as the same subject across evolutionary time with correlated measurements as in m-REDUCE, we treat them as four separate subjects that are correlated. Here we are making the assumption that the motif affects expression differently in each species, but within each species every motif that

²This thesis can be accessed at <http://proquest.umi.com/pqdweb?did=2099179251&sid=1&Fmt=2&VType=PQD&VInst=PROD&RQT=309&VName=PQD&TS=1336413805&clientId=18952>

functions as a TFBS will contribute the same additive effect to the expression of any gene with that motif in its promoter region. We constrain the regression coefficients by specifying the variance for the prior distributions of the intercept and the slopes to be proportional to the phylogenetic correlation matrix.

Formally, we make the following assumptions:

- Multivariate Normal (MVN) Distributions of Data and Parameters

$$\vec{Y}_g \sim MVN(\vec{\alpha} + \mathbf{X}_g \vec{\beta}, \sigma^2 \mathbf{I})$$

* $\vec{Y}_g$ is an $s \times 1$ vector of expression values (one for each of the s species) for gene g

* $\vec{\alpha}$ is an $s \times 1$ vector of intercepts (one for each of the s species) and is the same for all genes

* $\mathbf{X}_g$ is an $s \times s \times k$ design matrix of counts for the k motifs for gene g

* $\vec{\beta}$ is an $s \times k \times 1$ vector of k vectors, $\vec{\beta}_i$

* $\vec{\beta}_i$ is an $s \times 1$ vector of regression coefficients, (one for each of the s species) for motif i and is the same for all genes.

$$\vec{\alpha} \sim MVN(\mu_\alpha \vec{1}, \lambda_\alpha \mathbf{D})$$

$$\vec{\beta}_i \sim MVN(\mu_{\beta_i} \vec{1}, \lambda_{\beta_i} \mathbf{D}), \quad i \in \{1, 2, \dots, k\} \quad \forall \quad i$$

where $\mathbf{D}$ is the phylogenetic correlation matrix.

- Missing Data: $\vec{\alpha}, \vec{\beta}$
- Parameters: $\vec{\Theta} = \{\sigma^2, \mu_\alpha, \vec{\mu}_\beta, \lambda_\alpha, \vec{\lambda}_\beta\}$

Notice that all the $\vec{\beta}_i$ s are in one vector called $\vec{\beta}$ and all the μ_{β_i} s and λ_{β_i} s are in their own respective vectors in the parameters list to make the calculations easier when possible.

We used the Expectation Maximization (EM) algorithm [31] to calculate the maximum likelihood estimates (MLE) for the parameters in $\vec{\Theta}$ for each motif. We treated $\vec{\alpha}$ and $\vec{\beta}$ as missing data and $\vec{\Theta}$ as the list of parameters.

In this model the complete likelihood used for the EM algorithm was

$$\begin{aligned}
& L(\vec{\Theta}; \mathbf{X}, \vec{Y}, \vec{\alpha}, \vec{\beta}, \mathbf{D}) \\
&= Pr(\vec{Y} | \mathbf{X}, \vec{\alpha}, \vec{\beta}, \sigma^2) Pr(\vec{\alpha} | \mu_\alpha, \lambda_\alpha, \mathbf{D}) Pr(\vec{\beta} | \vec{\mu}_\beta, \vec{\lambda}_\beta, \mathbf{D}) \\
&\propto \prod_g \left\{ \left(\frac{1}{\sigma^2} \right)^{\frac{s}{2}} \exp \left[-\frac{(\vec{Y}_g - \vec{\alpha} - \mathbf{X}_g \vec{\beta})' (\vec{Y}_g - \vec{\alpha} - \mathbf{X}_g \vec{\beta})}{2\sigma^2} \right] \right\} \times \\
&\quad \left(\frac{1}{\lambda_\alpha^s |\mathbf{D}|} \right)^{\frac{1}{2}} \exp \left[-\frac{(\vec{\alpha} - \mu_\alpha \vec{1})' \mathbf{D}^{-1} (\vec{\alpha} - \mu_\alpha \vec{1})}{2\lambda_\alpha} \right] \times \\
&\quad \prod_i \left\{ \left(\frac{1}{\lambda_{\beta_i}^s |\mathbf{D}|} \right)^{\frac{1}{2}} \exp \left[-\frac{(\vec{\beta}_i - \mu_{\beta_i} \vec{1})' \mathbf{D}^{-1} (\vec{\beta}_i - \mu_{\beta_i} \vec{1})}{2\lambda_{\beta_i}} \right] \right\}.
\end{aligned}$$

For each motif, we used 10 random starting values for each of the parameters in $\vec{\Theta}$ and iteratively took the expectation of the log of the above likelihood, and then maximized the parameter estimates, until the difference between sequential parameter estimates were within a pre-specified value. Each parameter had its own stopping condition, as we found that some of the parameters had to be estimated more precisely than others to ensure consistent results. We ran the EM algorithm until all stopping conditions were met. Using the parameter estimates that gave the highest likelihood³ for that motif (out of the 10 random starts), we calculated the p-value, and iteratively chose the most significant motif to include in a forward selection manner. Since, in this model, $\vec{\beta}_i$ is not constant across all four species, we tested for the most significant motif by testing the mean of the prior distribution for $\vec{\beta}$ (i.e., μ_{β_i}) for every possible motif using the t-distribution to calculate a p-value for the two-sided test ($H_0: \mu_{\beta_i} = 0$ vs. $H_1: \mu_{\beta_i} \neq 0$). Significance is again measured at the 0.01 level with a Bonferroni correction for the number of motifs tested. This was implemented using R-2.6.1.

Since the parameters from this EM algorithm are MLEs and are from a large sample, we can consider them to be consistent and asymptotically efficient [32], and so can use normal theory to assign a p-value for significance testing. However, the variances of μ_{β_i} must first be calculated. Since this is a missing value problem and the estimated parameters are correlated, the variance was determined by extracting the element for the variance of μ_{β_i} in the inverted observed information matrix for the *incomplete* data. McLachlan and Krishnan [33] show how to calculate the observed information matrix for the incomplete data problem in terms of the complete log likelihood, $l(\vec{\Theta}; \mathbf{X}, \vec{Y}, \vec{\alpha}, \vec{\beta}, \mathbf{D})$.

Briefly, to find this observed information matrix, one first finds the conditional expectation of the *complete* information matrix given the observed data, and subtracts out the conditional expectation of the *missing* information matrix given the observed data. Specifically,

³Most of the time, the parameter estimates from all 10 random starting values were equivalent to the same number of significant figures as the stopping criteria, so there was very little difference among the likelihoods.

$$I(\vec{\Theta}; \mathbf{X}, \vec{Y}, \mathbf{D}) = \mathcal{I}_c(\vec{\Theta}; \mathbf{X}, \vec{Y}, \vec{\alpha}, \vec{\beta}, \mathbf{D}) - \mathcal{I}_m(\vec{\Theta}; \mathbf{X}, \vec{Y}, \vec{\alpha}, \vec{\beta}, \mathbf{D})$$

where

$I(\vec{\Theta}; \mathbf{X}, \vec{Y}, \mathbf{D})$ is the incomplete data information matrix,

$\mathcal{I}_c(\vec{\Theta}; \mathbf{X}, \vec{Y}, \vec{\alpha}, \vec{\beta}, \mathbf{D})$ is the conditional expectation of the complete information matrix given the observed data

$\mathcal{I}_m(\vec{\Theta}; \mathbf{X}, \vec{Y}, \vec{\alpha}, \vec{\beta}, \mathbf{D})$ is the conditional expectation of the missing information matrix given the observed data.

To calculate the above conditional expected complete and missing information matrices McLachlan and Krishnan [33] give the following formulas:

$$\mathcal{I}_c(\vec{\Theta}; \mathbf{X}, \vec{Y}, \vec{\alpha}, \vec{\beta}, \mathbf{D}) = -E \left[\frac{\partial^2 l(\vec{\Theta}; \mathbf{X}, \vec{Y}, \vec{\alpha}, \vec{\beta}, \mathbf{D})}{\partial \vec{\Theta}^2} \middle| \mathbf{X}, \vec{Y}, \mathbf{D} \right]$$

and

$$\mathcal{I}_m(\vec{\Theta}; \mathbf{X}, \vec{Y}, \vec{\alpha}, \vec{\beta}, \mathbf{D}) = E \left[\frac{\partial l(\vec{\Theta}; \mathbf{X}, \vec{Y}, \vec{\alpha}, \vec{\beta}, \mathbf{D})}{\partial \vec{\Theta}} \frac{\partial l(\vec{\Theta}; \mathbf{X}, \vec{Y}, \vec{\alpha}, \vec{\beta}, \mathbf{D})'}{\partial \vec{\Theta}} \middle| \mathbf{X}, \vec{Y}, \mathbf{D} \right].$$

In b-REDUCE, most of the μ_{β_j} s became highly significant ($-\log_{10}$ p-value > 30) after the third motif was found (see Section 4). We therefore selected the most significant motif at each iteration until we had a total of eight motifs to match the number of motifs found by m-REDUCE.

2.2.2 Single Species Model (s-REDUCE)—Both multiple-species models described above are based on REDUCE, the simple linear regression model of Bussemaker *et al.* [21] with modifications. To compare both multivariate models to a single-species model, we used the data from *S. cerevisiae* in a simple linear regression model (as in REDUCE) and incorporated the same modifications into the single-species model that we did into the multivariate models, so that the only difference was the number of species used. Specifically, our three modifications were:

1. Rather than calculate and use residuals as the response variable at each iteration as in REDUCE, we used a forward selection method where regression coefficients for all covariates were estimated at each iteration.

2. Rather than searching for motifs whose count is significantly correlated with the expression, we searched for motifs, using maximum likelihood estimation, that had a significant slope based on the t-distribution.
3. For the purpose of developing this model, we limited the motif length to 5 nucleotides, as the number of motifs grows exponentially by 4 with each nucleotide increase in length.

For the purposes of this paper, we refer to this single-species linear regression model as s-REDUCE.

2.3 Evaluations of Predicted Motifs

We used multiple lines of validation to assess the predictive accuracy of our models since there is no widely accepted standard for measuring performance of these types of algorithms. Here we briefly describe our evaluation methods, but for further details, see Siewert and Kechris [22]. To initially characterize the predicted motifs, we first counted the number of hits these motifs had to the canonical binding sites of the seven TFs listed in Harbison *et al.* [26] that were noted as being involved in the heat-shock stress response. Hits are defined either as an exact match or a match based on one shift (e.g., CTCGG is a shift for TCGGG). If a canonical motif had a degenerate position (fully degenerate positions were never allowed), then a predicted motif with any of the nucleotides at that position was still considered a hit, but a degenerate position was not allowed for a shift.

Secondly, the predicted motifs were evaluated based on the Kullback-Leibler information content, which is a measure of the conservation of a motif across species that accounts for the background probabilities of each base in the genome. We calculated the average information content per position across the four species for every motif found in the aligned promoter sequences of all the genes. Our assumption here is that if a predicted motif is functional, it will more likely be conserved across species, and thus have a higher information content compared to the background sequence.

We next generated a target gene list for each of our predicted motifs. This is a list of genes, each of which has an exact match to the predicted motif in its promoter region. These lists are compared to gene lists involved in the stressed condition based on two lines of evidence. First, as an independent source of stress responsive genes, we assembled a list of genes from the Gasch data set (see Section 2.1.3) that showed at least a two-fold increase of expression within 15 minutes of exposure to heat shock. We calculated the significance of the overlap between the motif target genes and the stress responsive Gasch genes using a Fisher's exact test. Here we are assuming that if the predicted motifs are functional, they should define target genes that are involved in the heat-shock stress response, and therefore, there should be a significant overlap of these genes and the heat-shock genes defined by the independent Gasch data.

Our second line of evidence using the target gene lists was to compare (for each motif) the binding strength of those motif target genes to the binding strengths of genes that did not contain that motif using the one-sided Kolmogorov-Smirnov test. This test distinguishes the difference between two distributions, and assigns a significance to that difference. A small p-

value indicates that the promoters containing the motif are distributed with a stronger binding strength than promoters that do not contain the motif, which (we are assuming) should be true if the predicted motifs are functional. For this evaluation, we used the eighteen data sets involved in heat shock from a comprehensive yeast ChIP-chip data set (see Section 2.1.3).

3 Results

The number of microarray expression data sets taken across multiple species has increased in the past 10 years. In conjunction with this type of data, new algorithms have been developed to extract this information for, among other applications, predicting TFBSs, *de novo*. In this work, we explored how best to constrain a multiple-species model to account for the correlations resulting from the phylogenetic relationships of these species. Starting with the multivariate, multiple linear regression model (m-REDUCE) which has a phylogenetic constraint on the variance of the error term, we altered the model to constrain the variance on the regression coefficients (b-REDUCE) rather than the error term. Imposing the phylogenetic constraint on the error term treats the four species as a single species with repeated, correlated expression measurements across evolutionary time. Imposing it on the regression coefficients treats the four species as four individual species that are correlated, each with their own separate expression measurement.

We first evaluated if both multivariate models improve predictive performance over using a single-species model regardless of where the phylogenetic constraint was imposed. We then compared the two multivariate models with each other to see how the placement of the phylogenetic constraint affected the predictive performance of the model.

For all model comparisons, we use four lines of evidence to evaluate predictive accuracy:

1. Motif hits to canonical TFBS sequences
2. Motif conservation across the species as measured by information content
3. Motif prediction of genes involved in the stress response as measured by the overlap of these genes with genes independently determined to be involved in the stress response
4. Binding strength of gene promoters predicted by the motifs to TFs involved in the stress response.

For more details, see Section 2.3.

3.1 Multiple Species versus Single Species

For the initial comparisons of multiple-species models (m-REDUCE and b-REDUCE) versus the single-species model (s-REDUCE), we used just the single 10 minute time point of the heat-shock stress response data set, because this response is a highly studied and well defined response and the 10 minute time point appeared to be the peak of the response based on a plot of gene expression across all genes for all time points (see Figure 2). We made the multiple-species versus single-species comparisons for all four species, but only show the results for *S. cerevisiae*. The results for the other three species were similar.

Table 1 lists the motifs found using each model. The numbers indicate the order in which the motifs were found. There are no motifs that were found by all three models. s-REDUCE and b-REDUCE share only one motif, s-REDUCE and m-REDUCE share four motifs, and b-REDUCE and m-REDUCE share only two motifs. The first motif found by both s-REDUCE and m-REDUCE was the same motif (AGGGG/CCCCT). While s-REDUCE found more motifs than either multivariate model (11 for s-REDUCE versus 8 for each multivariate model), the question remains concerning the quality of these motifs. Therefore, we characterized the motifs that were found by each model.

We initially evaluated the motifs found by the univariate model and both multivariate models using the seven TFs that were listed in Harbison *et al.* [26] as being involved with the heat-shock response. Table 2 shows the number of motifs that match or have a one-base shift to these canonical binding sites (No motifs had hits to Hsf1 or Yap1). Both m-REDUCE and b-REDUCE do as well, if not better than s-REDUCE at finding hits to the TFBSs involved in the heat-shock response. All three methods find motifs that have hits to these canonical binding sites, but s-REDUCE has four hits from the eleven motifs that it predicted, while each multivariate model has four hits from the eight motifs. Note that a single motif can have hits to different canonical binding sites. For example, although both m-REDUCE and b-REDUCE have four hits each to canonical binding sites, both find the motif ACCCC/GGGGT which is a match to Adr1 (with canonical site of MCYCC/GGRGK⁴) and a shift to Msn2 (with canonical site MAGGGG/CCCCTK). Therefore, three of the eight motifs found represent four hits to the canonical binding sites involved in heat shock. Note, also, that these hits can represent different motifs. For example, the hits to Adr1 (with canonical site of MCYCC/GGRGK) in m-REDUCE and b-REDUCE are both the motif ACCCC/GGGGT, while in s-REDUCE it is the motif ACTCC/GGAGT.

To distinguish between the motifs found by the multivariate and univariate models we further characterized them. We first evaluated their conservation across species by calculating the Kullback-Leibler information content of the motifs in aligned promoter sequences of the genes. Motifs that are more highly conserved are assumed to more likely be functional. Figure 3 shows dot and box plots of the average Kullback-Leibler information content per position per occurrence for the motifs found by s-REDUCE, b-REDUCE, and m-REDUCE. Figure 3a displays all motifs found by the three models, and Figures 3b and 3c display the motifs that are unique to s-REDUCE and either b-REDUCE or m-REDUCE, respectively. Note that in Figures 3b and 3c the motifs found by both models (s-REDUCE and either b-REDUCE or m-REDUCE) are not included. In contrast, part a) of these three figures includes all motifs. For example, the motif with an information content of 4.7 shown in the plots for s-REDUCE and m-REDUCE in Figure 3a is not shown in Figure 3c since both models found this particular motif. Comparing these types of plots serve to more clearly define the differences between two models. The plots in Figures 4 and 5 are similarly configured.

In all three plots, the median information content is higher for those motifs predicted by the multivariate models than by those predicted by the univariate model. More particularly, a

⁴Degenerate symbols: M = A or C, Y = C or T, R = A or G, K = G or T

greater percentage of the motifs found uniquely by each multivariate model had an average information content greater than 4.0 than those found uniquely by the univariate model. Specifically, 30% and 57% of the unique words found by s-REDUCE versus b-REDUCE and 29% and 100% of the unique words found by s-REDUCE versus m-REDUCE had information content greater than 4.0.

We next evaluated these motifs by examining how well they predicted genes that overlapped with genes independently defined to be involved in the heat-shock response. Figure 4 shows the significance of the overlap of these gene lists as measured by $-\log_{10}$ p-value of the Fisher's Exact Test. As in Figure 3, Figure 4a displays all motifs found by the three models, and Figures 4b and 4c display the motifs that are unique to s-REDUCE and either b-REDUCE or m-REDUCE, respectively. Again, all three plots show the improvement of the multivariate models over the univariate model. Looking more closely at the unique motifs that are shown in Figures 4b and 4c, 60% of the words uniquely found by s-REDUCE compared to 71% of the motifs uniquely found by b-REDUCE, and 86% of the words uniquely found by s-REDUCE compared to 100% of the motifs uniquely found by m-REDUCE had significant overlap (p-value < 0.05) with heat-shock responsive genes. Again, these results provide evidence for the advantages of multivariate prediction models over univariate models.

Lastly, we used the Kolmogorov-Smirnov test to evaluate how strongly TFs involved in the heat-shock response (as specified by Harbison *et al.* [26]) were bound to the promoters of genes predicted by the motifs compared to those genes not predicted by the motifs. Figure 5 shows dot and box plots of the significance of these tests expressed as $-\log_{10}$ p-value. We have shown the results from 12 of the eighteen data sets tested, as there were no significant differences in binding strength between those genes that were predicted and those that were not predicted for the other data sets for any of the motifs. Since there were 12 ChIP-chip data sets for each motif, each dot in the plots represents the median of the 12 results for that motif. These plots show that the motifs found by the two multivariate models tend to predict genes with promoters that are more tightly bound to the heat-shock TFs than those motifs found by the univariate model. Specifically, 60% of the motifs uniquely found by s-REDUCE compared to 57% of the motifs uniquely found by b-REDUCE, and 86% of the motifs uniquely found by s-REDUCE compared to 100% of the motifs uniquely found by m-REDUCE had median scores that were significant (p-value < 0.05).

Based on these three lines of evidence, we have shown that, although all three models find motifs that have hits to canonical TFBSs, the motifs found by the two multivariate models performed better than those motifs found by the univariate model. This would indicate that using information from multiple species to predict TFBSs, *de novo*, is an improvement over using just a single species. We next compared the two multivariate models to consider whether one model is better than the other.

3.2 b-REDUCE versus m-REDUCE

In this section, we compare the two multivariate models using the motifs uniquely found by each algorithm. Both are multivariate, multiple-linear regression models. The difference between them is the placement of the phylogenetic constraint. m-REDUCE constrains the

variance of the error term, while b-REDUCE constrains the variances of the regression coefficients. Although both models showed improvement over the single-species model, they did not predict the same set of motifs. Each algorithm resulted in eight motifs, but there was little overlap of these predicted motifs, as they only had two motifs in common (see Table 1). Table 2 shows that the motifs found by both models have four hits each to the seven known TFBSs involved in heat shock.

Again, as in the previous section, we distinguish between these two models by further characterizing the unique motifs found by each of these multivariate models. Figure 6a shows dot and box plots of the conservation across species of these motifs as measured by the Kullback-Leibler information content. Figure 6b shows plots of the significance of the overlap of the genes predicted by b-REDUCE and m-REDUCE with genes that are involved with the heat-shock response. And Figure 6c shows the binding strength of gene promoters predicted by these motifs to TFs involved with the heat-shock response. The dots in each set of plots represent the unique motifs found by b-REDUCE on the left and m-REDUCE on the right and the box plots overlaying the dot plots shows the median and the spread of these motifs. In these plots, it can be seen that the motifs found by m-REDUCE generally have higher conservation, predict genes that have greater overlap with heat-shock responsive genes, and predict gene promoters that are more tightly bound to heat-shock TFs than those motifs found by b-REDUCE.

We more carefully examined the specific results of these two models to investigate why m-REDUCE predicted higher scoring motifs than those found by b-REDUCE. The constraining matrix for both models was the same phylogenetic correlation matrix scaled by a free parameter (see Section 2.2.1). An inspection of the parameter estimates showed that the scaling parameter of the constraining matrix for the error term variance in m-REDUCE was estimated to be consistently in the range of 0.5. However, the scaling parameters of the constraining matrices for the variances of the regression coefficients, and in particular for the variance of the slope, was estimated to be in the range of $10^{-3} - 10^{-8}$ – essentially nullifying the phylogenetic correlation matrix. This very small parameter estimate would indicate that the slopes of the regression lines among the species are all the same, even though the model allows them to vary.

4 Discussion

Recently, an increasing number of microarray expression data sets taken across multiple species have been made available. However, there have been few studies on how to best take into account this type of correlated data in models. Using the heat-shock data set from Tirosh *et al.* [8] and an extension of a simple, linear regression model that predicts transcription factor binding sites, *de novo*, [21], we examined the two positions in a multivariate, multiple-linear regression model where the phylogenetic correlation can be included. m-REDUCE [22] imposes the phylogenetic correlation among the species on the error term. The method described in this paper (b-REDUCE) constrains the regression coefficients of this model using a Bayesian hierarchical approach. So that the constraining matrix was independent of the data, we obtained the correlation structure from published phylogenetic trees (see Section 2.2.1).

We found that both of these multivariate models improved predictive performance compared to the single-species model. Expanding the search space to multiple species may guard against both false positive and false negative results because it takes advantage of the information from other species and the phylogenetic relationships among them. For a motif to be found in our constrained multivariate models, that sequence needs to be relatively highly ranked in multiple species – not in just one species – but does not need to be the most highly ranked sequence in any of the species. In the univariate model, there is nothing to prevent either a spuriously top ranked sequence to be chosen as a putative TFBS (false positive) or a relatively highly ranked but not the most highly ranked sequence to be missed as a TFBS (false negative).

Comparing the two multiple-species algorithms with each other, we found that m-REDUCE performed better than b-REDUCE. In m-REDUCE, we required that the slope be the same for all genes *and* all species for each predicted TFBS. In b-REDUCE, however, we only required that the slope be the same for all genes for each predicted TFBS, but *not* for all species. That is, m-REDUCE treats the four species as the *same* subject across evolutionary time with correlated expression measurements, while b-REDUCE treats them as four *individual* subjects that are correlated. Our results showed that the scaling parameters (λ_α and λ_β) of the constraining covariance matrices for b-REDUCE were very small (see Section 3.2), essentially estimating zero variance of the regression coefficients among the species and eliminating the constraint of the phylogenetic matrix. This suggests that the assumption of varying slopes across species is unnecessary and may be a contributing factor to the poorer performance of b-REDUCE and that the phylogenetic relationship among the species is better incorporated in the error term of the model as in m-REDUCE.

Another consequence of this unmet model assumption was the high number of significant motifs generated. The extremely small scaling parameter estimates resulted in extremely small estimates of the variance and subsequent p-values for the slopes. Therefore, we chose the most significant motif at each iteration until there was a total of 8 motifs to correspond to the number of motifs found by m-REDUCE.

Both these models assume that if a binding site is functional, it will tend to be conserved across species. However, although both models rely on this conservation assumption, they use *unaligned* sequences as input data. The lack of an alignment requirement for the input sequence data is useful because, although TFBSs are conserved across species, they are not necessarily in the same location in all species [34]. Also, requiring aligned sequences as input data is less suitable for higher eukaryotes, where it may be difficult to align the long intergenic regions.

Furthermore, we used expression data from four species of yeast which span an evolutionary distance similar to that spanned between human and mouse [20], so these data are a good system to explore ways of imposing a phylogenetic constraint on multiple species. While our findings may not be applicable to models using data from more widely diverged species (e.g., worms to humans), they may be applicable to higher organisms that are less diverged (e.g., mouse to human).

Although the regression model that we used for a case study has, as input, sequence data, we recognize that better predictions may be achieved by alternative input data such as ChIP-Seq, or weighting the sequence data by external information such as methylation or nucleosome occupancy data. Our findings of the best placement of the correlation structure by constraining the error term rather than the regression coefficients might be relevant when using these types of data, as well.

In summary, we have presented a comparison of methods for incorporating correlated data from multiple species into a single prediction model. First, we found that using data from multiple species was more effective than using data from only a single species. More importantly, we found that the placement of the correlation constraint in a multivariate model seems to be important. Our results indicate that it is better to impose the correlation structure on the error term than on the regression coefficients.

5 Acknowledgments

This research was supported by a NIH/NLM Computational Bioscience Program training grant T15-LM00945 (ES) and a NIH/NIAAA grant K01-AA016922 (KK). We thank Itay Tirosh of the Naama Barkai lab for the expression data, Randy Wu, formerly of the Hao Li lab, for the sequence alignments, and John Siewert for computer run-time assistance.

References

- [1]. Ramsay G. DNA chips: state-of-the art. *Nature Biotechnology*. Jan; 1998 16(1):40–44.
- [2]. Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W, Paabo S. A neutral model of transcriptome evolution. *PLoS Biol*. 2004; 2(5):E132. [PubMed: 15138501]
- [3]. Lelandais G, Vincens P, Badel-Chagnon A, Vialette S, Jacq C, Hazout S. Comparing gene expression networks in a multi-dimensional space to extract similarities and differences between organisms. *Bioinformatics*. 2006; 22(11):1359–66. [PubMed: 16527831]
- [4]. Lemos B, Meiklejohn CD, Caceres M, Hartl DL. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. *Evolution*. 2005; 59(1):126–37. [PubMed: 15792233]
- [5]. Oleksiak MF, Churchill GA, Crawford DL. Variation in gene expression within and among natural populations. *Nat Genet*. 2002; 32(2):261–6. [PubMed: 12219088]
- [6]. Rifkin SA, Houle D, Kim J, White KP. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature*. 2005; 438(7065):220–3. [PubMed: 16281035]
- [7]. Eng KH, Bravo HC, Kele S. A phylogenetic mixture model for the evolution of gene expression. *Molecular Biology and Evolution*. 2009; 26(10):2363–72. [PubMed: 19602540]
- [8]. Tirosh I, Weinberger A, Carmi M, Barkai N. A genetic signature of interspecies variations in gene expression. *Nat Genet*. 2006; 38(7):830–4. [PubMed: 16783381]
- [9]. Bergmann S, Ihmels J, Barkai N. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol*. 2004; 2(1):E9. [PubMed: 14737187]
- [10]. Lefebvre C, Aude JC, Glemet E, Neri C. Balancing protein similarity and gene co-expression reveals new links between genetic conservation and developmental diversity in invertebrates. *Bioinformatics*. 2005; 21(8):1550–8. [PubMed: 15598835]
- [11]. McCarroll SA, Murphy CT, Zou S, Pletcher SD, Chin CS, Jan YN, Kenyon C, Bargmann CI, Li H. Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat Genet*. 2004; 36(2):197–204. [PubMed: 14730301]
- [12]. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science*. 2003; 302:249–255. [PubMed: 12934013]

- [13]. Chen C, Weirauch MT, Powell CC, Zamboni AC, Stuart JM. A search engine to identify pathway genes from expression data on multiple organisms. *BMC Syst Biol.* 2007; 1:20. [PubMed: 17477880]
- [14]. Lu Y, Rosenfeld R, Bar-Joseph Z. Identifying cycling genes by combining sequence homology and expression data. *Bioinformatics.* 2006; 22(14):e314–22. [PubMed: 16873488]
- [15]. van der Wath E, Moutsianas L, van der Wath R, Visagie A, Milanese L, Lio P. Grid methodology for identifying co-regulated genes and transcription factor binding sites. *IEEE Trans Nanobioscience.* 2007; 6(2):162–7. [PubMed: 17695752]
- [16]. Lu Y, Huggins P, Bar-Joseph Z. Cross species analysis of microarray expression data. *Bioinformatics.* 2009; 25(12):1476–1483. [PubMed: 19357096]
- [17]. Ward LD, Bussemaker HJ. Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics.* 2008; 24(13):i165–171. [PubMed: 18586710]
- [18]. Kechris K, Li H. c-REDUCE: Incorporating sequence conservation to detect motifs that correlate with expression. *BMC Bioinformatics.* 2008; 9(1):506. [PubMed: 19040743]
- [19]. Liu Y, Liu XS, Wei L, Altman RB, Batzoglou S. Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res.* 2004; 14(3):451–8. [PubMed: 14993210]
- [20]. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature.* 2003; 423(6937):241–54. [PubMed: 12748633]
- [21]. Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. *Nat Genet.* 2001; 27(2):167–71. [PubMed: 11175784]
- [22]. Siewert EA, Kechris KJ. Prediction of motifs based on a repeated-measures model for integrating cross-species sequence and expression data. *Statistical Applications in Genetics and Molecular Biology.* 2009; 8(1)
- [23]. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science.* 2003; 301(5629):71–6. [PubMed: 12775844]
- [24]. Thompson J, Higgins D, Gibson T. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994; 22(22):4673–80. [PubMed: 7984417]
- [25]. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell.* 2000; 11(12):4241–4257. [PubMed: 11102521]
- [26]. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature.* 2004; 431(7004):99–104. [PubMed: 15343339]
- [27]. Rokas A, Williams BL, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature.* 2003; 425(6960):798–804. [PubMed: 14574403]
- [28]. Kuramae EE, Robert V, Snel B, Weiss M, Boekhout T. Phylogenomics reveal a robust fungal tree of life. *FEMS Yeast Research.* 2006; 6(8):1213–1220. [PubMed: 17156018]
- [29]. Holland BR, Huber KT, Moulton V, Lockhart PJ. Using consensus networks to visualize contradictory evidence for species phylogeny. *Mol Biol Evol.* 2004; 21(7):1459–61. [PubMed: 15084681]
- [30]. Speed, T. *Statistical Analysis of Gene Expression Microarray Data.* Interdisciplinary Statistics. Chapman and Hall/CRC; Danvers, MA: 2003.
- [31]. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological).* 1977; 39(1):1–38.
- [32]. Casella, G.; Berger, RL. *Statistical Inference.* Duxbury Advanced Series. 2nd. Duxbury; Pacific Grove, CA: 2002.
- [33]. McLachlan, JG.; Krishnan, T. *The EM Algorithm and Extensions.* Wiley Series in Probability and Statistics. 2nd. John Wiley & Sons, Inc.; Hoboken, NJ: 2008.

- [34]. Berman B, Pfeiffer B, Lavery T, Salzberg S, Rubin G, Eisen M, Celniker S. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biology*. 2004; 5(9)

Figure 1.

Phylogenetic Tree Relating Five Yeast Species. The topology of the the phylogenetic tree for the five species is published in [27–29]. The branch lengths (solid lines) of the phylogenetic tree for the four species is published in [20] . The published branch lengths did not include that of *S. kudriavzevii* (dashed line). We therefore used an average branch length between *S. mikatae* and *S. bayanus* for this branch length as noted in parentheses. The branch lengths shown represent the number of substitutions per site in intergenic sequence across the entire genome.

Figure 2. Gene Expression across Time. The number of genes with a two-fold increase (positive axis) or decrease (negative axis) of expression in *S. cerevisiae* are plotted for each time point. The number of genes responding to the heat shock was highest at the 10 minute time point.

Figure 3.

Comparison of Multivariate versus Univariate Models - Information Content. Each plot displays the average Kulback-Leibler information content per position per occurrence for each motif found by s-REDUCE, b-REDUCE, and m-REDUCE. The dots represent the individual motifs, and the overlaying box plots show the median and the spread for the motifs. The dotted line at 3.3 is the background information content as measured by the average information content per position across all promoter sequences of all genes. The results shown for s-REDUCE are the *S. cerevisiae* results. a) All motifs found by each model b) Motifs found uniquely by s-REDUCE and b-REDUCE c) Motifs found uniquely by s-REDUCE and m-REDUCE

Figure 4.

Comparison of Multivariate versus Univariate Models - Overlap of Target Genes with Heat-shock Genes. Each plot displays the $-\log_{10}$ p-value of the Fisher's Exact Test for the overlap of genes responsive to heat shock and target genes predicted by motifs found by s-REDUCE, b-REDUCE, and m-REDUCE. The dots represent the individual motifs, and the overlaying box plots show the median and the spread for the motifs. The dotted line at 1.3 marks the significance at the 0.05 level. The results shown for s-REDUCE are the *S. cerevisiae* results.

a) All motifs found by each model b) Motifs found uniquely by s-REDUCE and b-REDUCE
c) Motifs found uniquely by s-REDUCE and m-REDUCE

Figure 5.

Comparison of Multivariate versus Univariate Models - Binding Strength of Target Gene Promoters to Heat-shock Transcription Factors. Each plot displays the $-\log_{10}$ p-value of the Kolmogorov-Smirnov tests for the difference in binding strengths of gene promoters predicted by each motif compared to gene promoters not predicted by each motif found by s-REDUCE, b-REDUCE, and m-REDUCE. The dots represent the median $-\log_{10}$ p-value of the 12 ChIP-chip data sets for each motif, and the overlaying box plots show the median and the spread for the motifs. The dotted line at 1.3 marks the significance at the 0.05 level. A significant dot represents a motif in which at least half of the 12 data sets were significant. The results shown for s-REDUCE are the *S. cerevisiae* results. a) All motifs found by each model b) Motifs found uniquely by s-REDUCE and b-REDUCE c) Motifs found uniquely by s-REDUCE and m-REDUCE

Figure 6.

Comparison of b-REDUCE versus m-REDUCE. The plots display the results for the motifs found uniquely by each multivariate model. The dots represent the individual motifs, and the overlaying box plots show the median and the spread for the motifs. a) This plot displays the average Kullback-Leibler information content per position per occurrence for each motif found uniquely by b-REDUCE and m-REDUCE. b) This plot displays the $-\log_{10}$ p-value of the Fisher's Exact Test for the overlap of genes responsive to heat shock and target genes predicted by motifs uniquely found by b-REDUCE and m-REDUCE. c) This plot displays the $-\log_{10}$ p-value of the Kolmogorov-Smirnov tests for the difference in binding strengths of gene promoters predicted by each motif compared to gene promoters not predicted by each motif uniquely found by b-REDUCE and m-REDUCE. In this plot, the dots represent the median p-value of the 12 ChIP-chip data sets for each motif.

Table 1

Motifs Found by s-REDUCE, b-REDUCE, and m-REDUCE Models. The first 11 motifs in column 1 are sorted in the order they were found by s-REDUCE, while the last 9 motifs were unique to b-REDUCE and/or m-REDUCE. The numbers in columns 2 – 4 indicate the order of discovery of the motifs for each model. The results shown for s-REDUCE are the *S. cerevisiae* results.

Motif	s-REDUCE	b-REDUCE	m-REDUCE
AGGGG/CCCCT	1	-	1
ATCGC/GCGAT	2	5	-
GAAAA/TTTTTC	3	-	5
AAGGG/CCCTT	4	-	-
ACTCC/GGAGT	5	-	-
ATGAG/CTCAT	6	-	-
ATAGG/CCTAT	7	-	-
GTCCA/TGGAC	8	-	-
AAATT/AATTT	9	-	2
AACAT/ATGTT	10	-	-
CATCG/CGATG	11	-	7
CTCCC/GGGAG	-	1	-
TGAAA/TTTCA	-	2	-
ACCCC/GGGGT	-	3	4
AAAAA/TTTTT	-	4	-
AGGGA/TCCCT	-	6	6
ATAAA/TTTAT	-	7	-
CCTCG/CGAGG	-	8	-
GCCAC/GTGGC	-	-	3
CACCC/GGGTG	-	-	8

Table 2

Hits to Canonical TFBS for Motifs Found in s-REDUCE, b-REDUCE, and m-REDUCE. The transcription factors that have motif hits to their canonical TFBS are listed in the first column. The number of predicted motifs having hits to each canonical TFBS is listed for each algorithm in columns 2-4. Hits are either defined as an exact match or a match based on one shift (e.g., CTCGG matches TCGGG). If a canonical motif had a degenerate position (excluding fully degenerate positions), then a predicted motif with any of the nucleotides at that position was still considered a hit, but a degenerate position was not allowed for a shift. For example, the predicted motif, AAGGG, would match the canonical motif MAGGGG since M is a degenerate symbol for A or C, but AAAGG would not be counted as a hit, since the shift would include the degenerate symbol, M. The results shown for s-REDUCE are the *S. cerevisiae* results.

Transcription Factor	s-REDUCE	b-REDUCE	m-REDUCE
Adr1	1	1	1
Gat1	-	1	-
Msn2	2	1	2
Skn7	1	-	1
Xbp1	-	1	-