Final report

Quantifying how Cis-regulatory elements predict multi-dimensional mRNA expression programs

Student: Hanqin Du
Supervisor: Wallace Edward
Co-Supervisor: Haynes Samuel

4th Year Project Report
Artificial Intelligence and Computer Science
School of Informatics
University of Edinburgh
2020

Abstract

Gene expression is a process of synthesising genetic products with various functions such as proteins. To survive, cells must be able to respond to environmental stress by adjusting the level of gene expression. A large number of studies have made efforts to reveal the complicated mechanism of the regulation of gene expression. However, most of these studies are focusing on finding and analysing cis-regulatory elements involved in pre-transcriptional regulation, such as promoters and enhancers. Many reports indicate that post-transcriptional regulation also plays an important role in gene expression regulation. Post-transcriptional regulation refers to regulation at the post-transcriptional level (mRNA). It includes the regulation of multiple aspects, such as mRNA modification, transfer, stability regulation, and degradation, etc. In this study, we applied group lasso as a multi-task learning method to quantify the regulatory effect of the cis-regulatory elements located on the 3'UTR of mRNA. Base on the model, we suggest a series of approaches for accessing the significance of each factor and comparing the regulatory effect between factors by analysing the regression coefficient from group lasso. At last, we proposed the cis-regulatory elements worth experimentally verifying and discuss the benefit and problem of our approaches.

Please read `hanqin_summary.Rmd` for core source code and `hanqin_summary.html` for its output.

Acknowledgements

I would like to thank my supervisor, Dr Edward Wallace and Samuel Haynes for their patient guidance, detailed suggestion and help in building my knowledge in biology.

I would also like to thank Abhishek Jain for sharing his honours project report, codes and data with me.

Table of Contents

Chapter 1

Introduction

mRNA (Messenger Ribonucleic acid) is a copy of part of the DNA sequence and is used to deliver genetic information from the nucleus to the cytoplasm where proteins are made. There are four types of nitrogenous bases found in mRNA: 1.) **A**-adenine, 2.) **C**-cytosine, 3.) **G**-guanine, 4.) **U**-uracil, which are transcribed from DNA bases 1.) **A**-adenine, 2.) **C**-cytosine, 3.) **G**-guanine, and 4.) **T**-thymine respectively. Therefore, the sequence of mRNA can be expressed as a string constructed with characters **A,C,G,U**. During the translation process, when mRNA is 'translated' to protein, nucleotide triplets will be mapped (surjective) and translated to one of the 20 amino acids. However, not all the sequence of mRNA is translated into protein. There are untranslated regions (UTR) sit at both ends of mRNA. Base on their chemical structure, they are named 5'UTR and 3'UTR. These noncoding regions have been proven to have a strong regulation effect on gene expression [1].

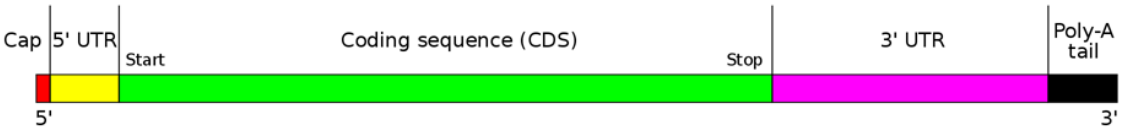


Figure 1-1. *The structure of a typical protein coding mRNA including the untranslated regions (UTRs). figure adapted from [29]*

The regulation of gene expression ,which affects the protein level and the function of a cell, is the main way that cells respond to environmental changes. The level of a particular protein at a certain time depends on the balance between that protein's synthetic and degradative biochemical pathways. For the synthetic pathways, the production of protein starts when it is transcribed from DNA to mRNA and continues with translation from mRNA to protein (Figure 1-2). Thus, controling these processes plays a critical role in determining which proteins are present in a cell and the amount of protein that is present [1].

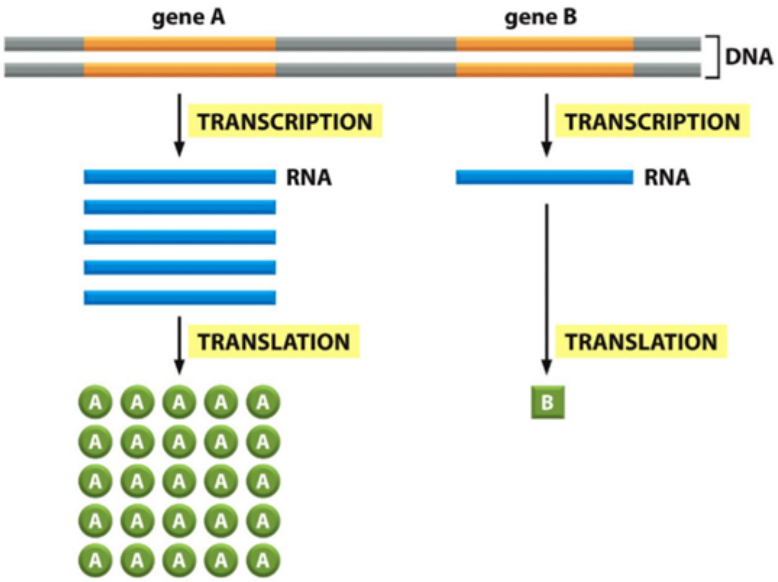


Figure 1-2. *the more mRNAs there are, the more proteins will be synthesized. Figure adapted from [1].*

Cis-regulatory elements(CREs) are sequences located in or close to a gene in the primary structure that have regulatory effect on that gene. One way CREs regulate gene is by providing binding sites for RNA-binding proteins (RPBs), which can be found across the whole message but are more usually located in the 5'- or 3'-untranslated regions of mature mRNA sequences and normally have an important regulatory effect on the mRNA [2-4]. In fact, many studies have pointed out that the sequence of 3′UTR is significantly related to the stability of the associated mRNA [5-8]. Take the Puf proteins' family in yeast as an example: the Puf proteins can bind to 3'UTR sequences encompassing UGUR tetranucleotide motif and thus repress gene expression by affecting mRNA translation or stability [9]. Moreover, Randi J. Ulbricht et. al's study also suggests that TIF1Puf1p and Puf5p can even act in tandem and recognize two UGUA binding sites within the TIF1 3' UTR, thus controlling its decay rate [10].

Identyfing the cis-regulatory elements (CREs) and quantifying their regulatory effect has been a popular research topic. Various methods have been developed and applied to locate cis-regulatory elements. Candidate CREs can be selected by immunoprecipitation of TF or histone post-translationally modified chromatin, identifying "open" chromatin regions of nucleosomes, or high-throughput functional screening based on sequencing [11-16]. Computational methods like comparative genomics, clustering and supervised machine-learning on an existing dataset could also be used to predict candidate CREs [13,14,17]. The accumulation of research over the years has provided sufficient conditions for the quantitative analysis of CRE effects.

However, a majority of the past research that aims to quantify the effect of CRE focuses on promoters and enhancers which regulate transcription while ignoring the effect of UTR on mRNA in post-transcriptional regulation. Statistic and machine learning methods , including K-means clustering, Bayesian networks model, and In silico genome-context analysis have been applied to select cis-regulatory elements from the DNA sequence that are likely to be enhancers or promoters and investigate how they affect expression level under environmental stress [14, 18].

Moreover, the environmental condition chosen for these quantitative studies of CRE regulatory effects is very limited. The main reason for considering multiple environmental stresses is that an RBP may affect the gene expression in a completely different way under two distinct environmental stress conditions. Alfredo Castello et al. provided a good example: Iron Regulatory Protein 1 (IRP1) is an RBP that balances the free iron concentration. Under low intracellular concentrations of iron, IRP1 can either bind to 5'UTR of mRNA and repress its translation, or binds to 3'UTR to increase the stability of transferring the receptor mRNA. However, when the intracellular iron concentration rises, IRP1 becomes active as cytosolic aconitase, which catalyzing the interconversion between citrate and isocitrate (Figure 1-3) [19].

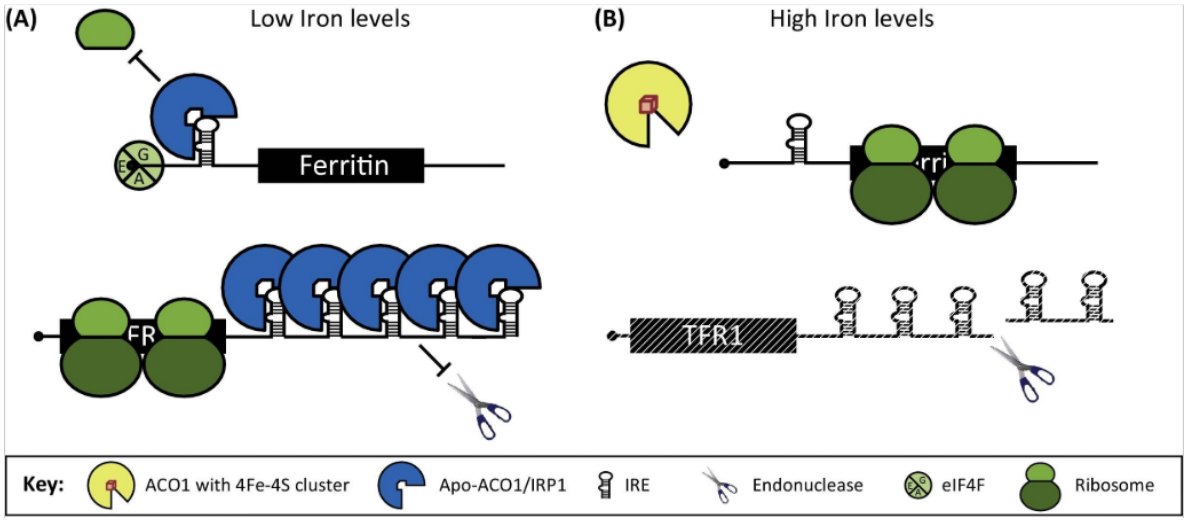


Figure 1-3. *IRP1 represses translation under low iron level and active as cytosolic aconitase under high Iron levels. Figure adapted from[19].*

Therefore, analyzing the regulatory effect of CREs under a series of different environmental stresses is important. Although a linear model can provide easy-to-understand regression coefficients for quantifying regulatory effects of each factor, it can only predict how the gene expression level change under single environmental condition. Because cells respond to different environmental stresses differently, we may need multiple linear models for multiple environmental conditions. Furthermore, a simple linear model may have many coefficients non-zero yet insignificant, thus weak false positives. The **group lasso**, introduced by Yuan et. al, provides a viable solution for training multiple models at once and comparing the significance of their coefficients. The group lasso is proposed to find important explanatory factor groups in regression models. Although it is more computationally expensive than the other two approaches mentioned in the same paper, it is widely used due to its excellent performance and stability (usually reaches a reasonable convergence tolerance within a few iterations). Note that the other two methods are `Group least angle regression selection` and `Group non-negative garrotte`. The group lasso penalty is computed by penalizing coefficients from each group individually by a symmetrical d x d positive definite kernel matrix K_j, and then summing up [20]:

$$\lambda \sum_{j=1}^{J} ||\beta_j||_{K_j}$$

In this project, we collected data from previous studies of gene expression, transcripts and motifs, and supplements from online databases and constructed a design matrix (Chapter 2). A previous honours student Abhishek Jain have applied penalised regression to estimate the role of these motifs in the one-dimensional outputs of RNA half-life, but did not address stress responses [28]. In Chapter 3, frequency analysis, correlation coefficient, and linear regression were used to investigate the dataset and evaluate whether the existing data meets suitable conditions to apply the group lasso. In Chapter 4, we first applied the kNN imputation recommended by Troyanskaya et al. [25] to handle the missing gene expression data from Gasch et al to obtain a complete design matrix without missing values. Then we selected hyperparameter for group lasso by cross-validation and then fitted the model. Chapter 5 describes a series of analyses we carried out on the regression coefficient provided by the group lasso model where plots are generated with package `ggplot2` [27]. By considering the transcript frequency, the L2 norm of the regression coefficient, and the relationship between each sub-model at the same time, we obtained some conclusions that may provide guidance for future experiments and selected the six significant motifs which perform differently under different environmental stresses. To test the reliability of the six motif's regression coefficients, we fitted linear models with these motifs and checked their coefficients and standard error. Finally, in Chapter 6, by summarizing our results and conclusions, we propose motifs worth experimentally verifying and discussed the benefit and disadvantage of our approaches.

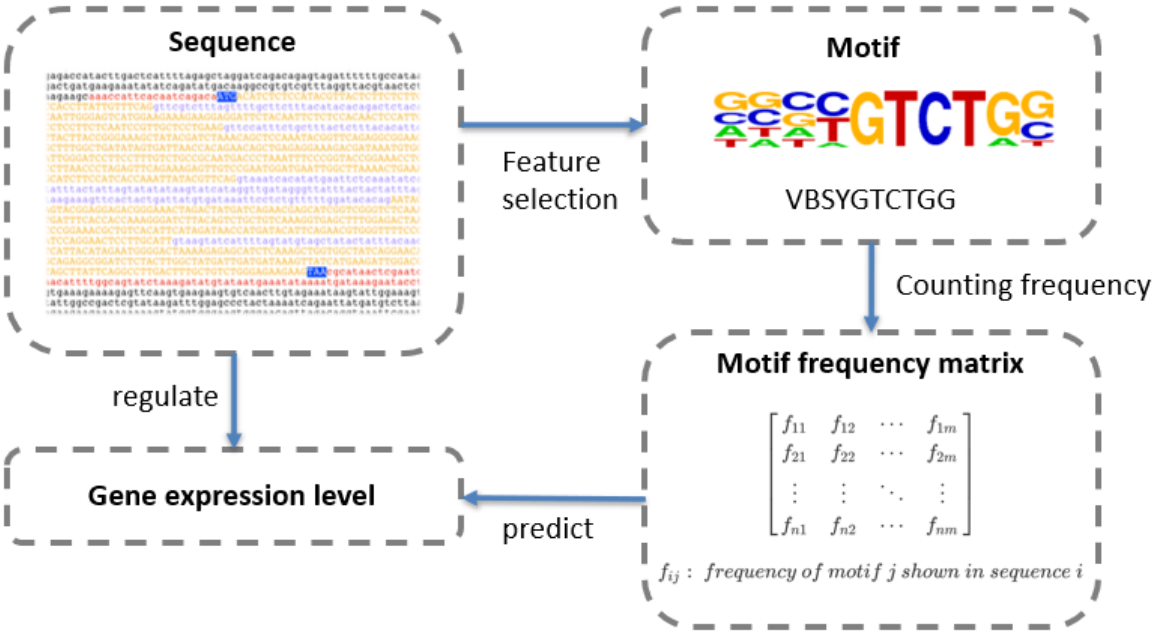


Figure 1-4. *the basic idea of our model. By scanning the sequence and counting the motif, we obtained the motif frequency matrix where each row represents a different gene and each column represents the frequency of the motif of that gene. Assuming we know how each motif responds to different environmental stress, This matrix could be used to explain how gene expression levels would change under specific environmental stress.*

Chapter 2

Prepare Data

2.1 Expression level change data from Gasch's study

In this project, we imported the data describes gene expression level in various environment condition from Gasch's study where DNA microarrays were applied to measure the relative abundance of mRNA (relative transcript level) before and after a series of environmental stresses. The dataset consists of 6152 rows and 176 columns where each row represents a gene of yeast. The first 3 columns provide the details of the gene and the other 173 columns describe the normalized, background-corrected log2 values expression level change under 173 different environmental stress [21].

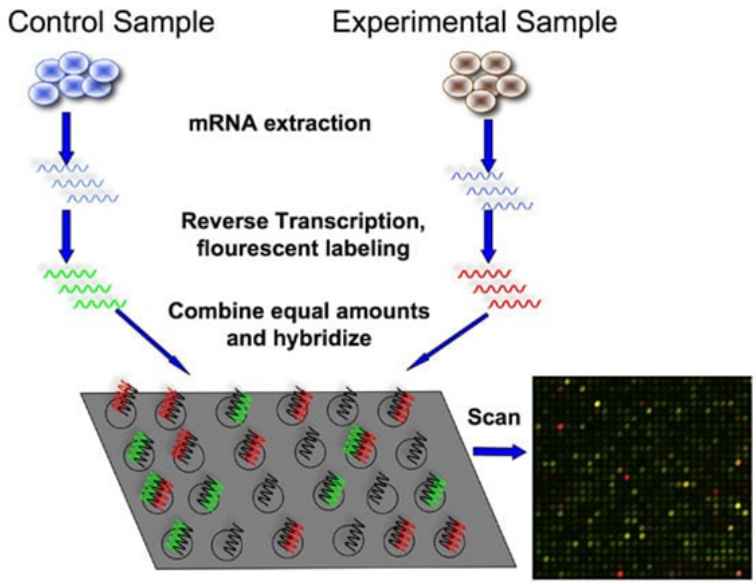


Figure 2-1. *The general process of microarrays. The flow figure describes how to compare the relative quantity of mRNA (relative transcript level) of specific genes in Control Sample and Experimental Sample through microarrays. First, mRNA is extracted from the cells and add the stained cDNA. After the cDNA combined with the mRNA, the original mRNAs are removed to obtain the reverse complementary sequences of the specific sequences on the original mRNA. Then these reverse complementary sequences with different colours are mixed and placed on the microarray chip. There are thousands of dot-shaped areas on the chip where each of them contained different prepared sequences. After the reverse complementary sequences binding with the sequences on the chips, the chip is washed and scanned in two colours which indicate the source of the sequences respectively. The relative density of each colour shows the relative transcript level of each sample. (Bitesizebio, Introduction to DNA Microarrays). Figure adapted from [30].*

Figure 2-2 shows the names of the first few columns of the gene expression profile. The last five columns of the table represent the log2 ratio between the relative transcript level at 5, 10, 15, 20, 30, 40, 60 and 80 minutes after heat shock and the relative transcript level before heat shock. Time slices between different types of environmental stresses are not aligned. For example, the dataset contains relative transcript levels at 5, 10, 15, 20, 30, 40, 60, and 80 minutes after `heat shock from 25°C to 37°C` while for `menadione exposure`, relative transcript levels of 10, 20, 30, 40, 50, 80, 120, and 160 minutes were recorded. In nitrogen exposure, time is even recorded in hours and days. This increase the difficulty of the cross-comparison across the regression coefficient between different types of environmental stresses.

	geneName	hs_05min_hs-1	hs_10min_hs-1	hs_15min_hs-1	hs_20min_hs-1	hs_30min_hs-1	hs_40min_hs-1	hs_60min_hs-1	hs_80min_hs-1
1	YAL001C	1.53	-0.06	0.58	0.52	0.42	0.16	0.79	*NA*
2	YAL002W	-0.01	-0.30	0.23	0.01	-0.15	0.45	-0.04	0.14
3	YAL003W	0.15	-0.07	-0.25	-0.30	-1.12	-0.67	-0.15	-0.43
4	YAL004W	0.24	0.76	0.20	0.34	0.11	0.07	0.01	0.36
5	YAL005C	2.85	3.34	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*

Figure 2-2. *A preview of the expression level profile.*

2.2 The 69 candidate sequence motifs on RNA untranslated region are selected from 3 different studies:

A motif indicates a sequence pattern which is considered to have a biological significance. We imported sequence motifs that believed to contain CRE from three studies and expressed them as regular expression under IUPAC rules. For example, motif `A[CG]T[ACGT]` can be written as `ASTN`. Note that during RNA transcription, **T**-thymine is usually transcribed into **U**-uracil, thus **T** is synonymous to **U** in the context of motif names.

53 candidate sequence motifs came from the study of Shalgi et al (2005), where motifs are derived by analyzing the exhaustively enumerating all k-mers(k = {8,9,10,11,12}) and looking for over-represented motifs with extreme half-life value. 515 significant k-mers were selected with ranksum test and clustered by ClustalW which resulted in 51 clusters of motifs. Additionally, two more motifs were found by grouping genes with extreme half-life and ran Gibbs sampler [22].

14 candidate sequence motifs came from the study of Hogan et al. (2008). Two related computational methods were applied to identify candidate binding sites of 40 out of the more than 500 known and predicted RBPs in S. cerevisiae: (1)“finding informative regulatory elements” (FIRE) and (2)“relative filtering by nucleotide enrichment” (REFINE). The former searches for motifs with informative patterns of enrichment. The other one identifies all hexamers that are significantly enriched in untranslated regions, filters out regions of target sequences that are relatively devoid of such hexamers, and then applies the "multiple expectation maximization for motif elicitation" (MEME) motif-finding algorithm. As a result, 14 motifs that are likely to be the binding sites of 16 RBP are found [23].

The rest 4 were from the study of Cheng (2017): four mRNA-stability related motifs in the 3' UTR were found by De novo motif searching and their reliability and effect have been examined by linear mixed effect model, Fisher test P-value corrected with Benjamini–Hochberg, Wilcoxon rank-sum test and multivariate linear regression [24].

The 69 sequence motifs are listed in the Appendix

2.3 3'UTR data gathered from previous study, online database and lab

The majority of 3'UTR data came from the research of Cheng et al. [24]. Annotated transcript sequences data of 4284 genes out of 6152 genes were downloaded from the GitHub sites introduced in their paper.

For the rest of the 1868 genes, since there is no reliable 3'UTR annotation in most of the online database, we applied downstream sequence with fix-length (120 nucleotides) as 3'UTR. The length of 120 nucleotides is considered as a sensible number close to the median length. Therefore, we downloaded 1320 downstream sequences fungiDB and imported 503 downstream sequences from Samuel's lab dataset.

To estimate the bias of applying 120-nucleotides downstream sequence as 3'UTR sequence. We compared the frequency of each motif per gene between 120-nucleotides downstream sequences and actual 3'UTR sequences of the 4284 genes. As shown in Figure 2-3, the similarity between the motifs' frequency distribution from both datasets ensured the fit-length downstream sequence could be applied as an alternative of 3'UTR sequence. A drawback of such approach is that when applying the 120 downstream sequences, a small part of the motifs which sit after 120 downstream nucleotides is missed. This may lead to an overestimation of the regulatory effect of some motifs. While considering that importing downstream sequences could increase 40% in the sample size, we believed the benefit of applying 120 downstream sequences overwhelm its drawback.

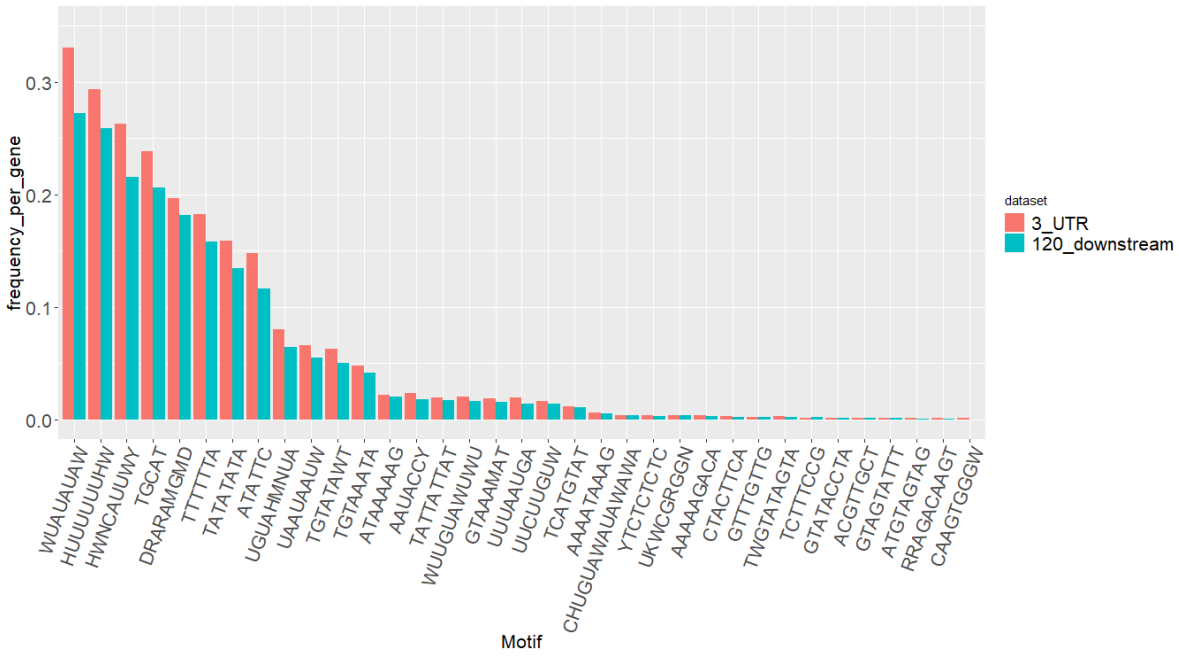


Figure 2-3. *The motifs frequency per gene of actual 3'UTR and 120 downstream sequence*

	genename	UTR3_seq
1	YAL002W	CATTTCTAAATATTTAATACAACTTTGGTTACATAAAAGTAAAATTTATACACCTC
2	YAL003W	AAGGCTTTTTTATAAACTTTTTATAATTAACATTAAAGCAAAAACAACATTGTAAAGATTAACAAATAAATGAAAAAAACAACGAAATAACTTAGGTTTTAGGCTAAAAAAAACAGAAGGAATTTTGAACGATAAACTTTTCGACTG...
3	YAL005C	GCCAATTGGTGCGGCAATTGATAATAACGAAAATGTCTTTTAATGATCTGGGTATAATGAGGAATTTTCCGAACGTTTTTACTTTATATATATATACATGTAACATATATTCTATACGCTATAGAGAAAGGAAATTTTTCAATT
4	YAL007C	GAACTTTTCAATCTACGAAAAATATATGTCCGCAATATAGAACACAATTAGGTTTATATTCGACGTGATTTTTTTTTCTTCCTTAGCCCTATGTATATTTACTGTATAGGATAAATGAAATACCAAAAATAAAAAAGTATAAAACG
5	YAL008W	GCAAGACAAATGACCAGATATAAACGAGGGTTATATTCTTTCGTTTTATACTTTTTTATTTTTGGTATTTCATTTATCCTATACAGTAAATATACATAGGGCTAAGGAAGAAAAAAAAATCACGTCG

Figure 2-4. *A preview of the 3'UTR data.*

2.4 Construct design matrix from gene expression profile and motifs frequency profile

To construct the design matrix, we required both the gene expression level data from Gasch et al. and the motif frequency profile which indicates the frequency of each of 69 motifs on the 3'UTR sequence of each gene. Therefore we begin with scanning candidate motifs on 3'UTR sequences to construct the motif frequency matrix (Figure 2-6). Where each row represents a gene, and each column represents the frequency of a specific motif on the 3'UTR of that gene. Since we focus only on post-transcriptional control but not the overall regulation of gene expression, we did not consider the reverse complement sequences when scanning the 3'UTR sequences (Figure 2-4). This is one of the main differences between our study and previous relavant study. As shown in Figure 2-5, DNA is a regular double helix structure with two nucleotide sequences where the bases on them matching each other one by one (C=G, A≡T). The regulatory factor can target the sequence on both of the base chains as a binding site to regulate the gene expression. Therefore, it is important to consider both sequences when studying promoters and enhancers. In our project, since we only focused on the single strand produced by transcribed DNA - mRNA. Thus only the sequence on the mRNA is considered when counting motifs.

Figure 2-5. *From DNA to mRNA. Figure adapted from [31].*

	geneName	ATATTC	TGCAT	TGTAAATA	TTTTTTA	CHUGUAWAUAWAWA	UGUAHMNUA	WUUGUAWUWU	UAAUAAUW	AKUCAUUCCUU	WUAUAUAW
1	C5_05510C_A	0	0	0	0	0	0	0	0	0	0
2	YAL001C	1	0	0	0	0	0	0	0	0	0
3	YAL002W	0	0	0	0	0	0	0	0	0	0
4	YAL003W	0	0	0	1	0	0	0	0	0	0
5	YAL005C	1	0	0	0	0	1	0	0	0	1

Figure 2-6. *A preview of the motif frequency profile.*

After obtaining the frequency matrix, we checked the total frequency of each motif appearing on 3'UTR (Table 2-7, Figure 2-8). We noticed that seven motifs have never been observed on any of the 3'UTR sequences. According to the description of the paper providing these motifs, they may only exist on 5'UTR. Therefore we removed them so that only 61 motifs left.

Frequency	number of motifs (n)
n = 0	7
0 < n < 6	8
5 < n < 21	7
20 < n < 51	25
n > 50	22

Table 2-7. *frequency distribution of motifs*

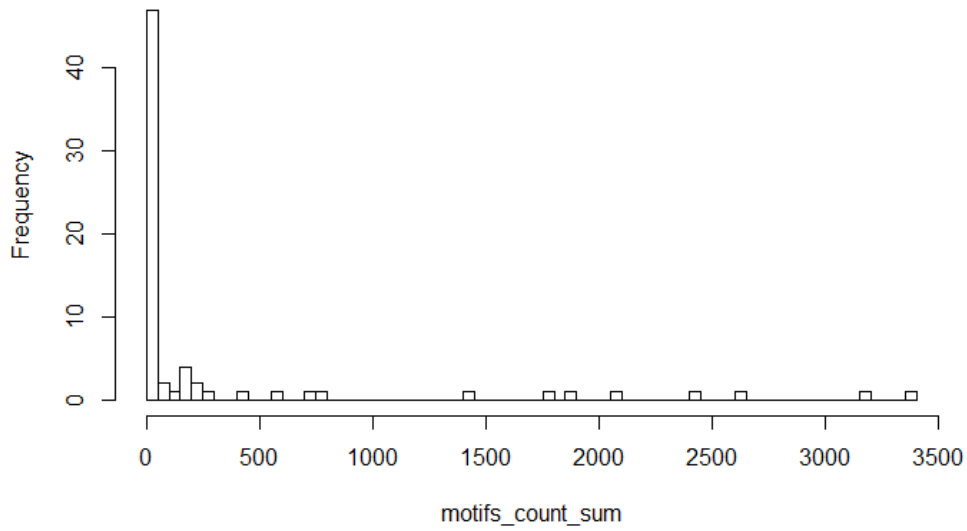


Figure 2-8. *The histogram of regquency distribution of motfs*

Finally, we inner joined the gene expression level (Figure 2-2) profile with the motif frequency matrix (Figure 2-6) by gene name to construct the design matrix.

Chapter 3

Justify group Lasso

We applied the group lasso for three reasons. First, it allows us to train multiple models together where each model explains how the motifs relate to the expression level change under different environmental stresses. Second, it helps us selecting significant motifs that show strong regulatory effect during environmental stress. And last, the lasso term could penalize the model to prevent overfitting.

In order to justify whether it is appropriate to apply group lasso on current dataset, we investigated: (1) if there is potential linear relation between motifs and the change of gene expression level; (2) if there are significant motifs with strong regulatory effect under certain type of environmental stress; (3) if the regulatory effect of motifs changes under different types of environmental stresses.

3.1 4 heat-shock-relevant motifs are selected with high correlation coefficient

To investigate how the log2 change of gene expression level depends linearly on the motifs frequency, we calculate the Pearson correlation coefficient (PCC) between them. Pearson correlation coefficient is a measure of the linear correlation between two variables X and Y:

$$p_{x,y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

We started by picking a group of environmental conditions about heat shock from Various Temperatures to 37°C. Four motifs showed a much higher correlation coefficient rather than the others, which, indicated the linear relationship between these motifs and expression level. One point worth mentioning is that all these motifs have been mensioned in Abhishek's report where `TGTATAWT` was expected to be positively associated with RNA stability and the rest three were expected to be negatively associated with the RNA stability [28].

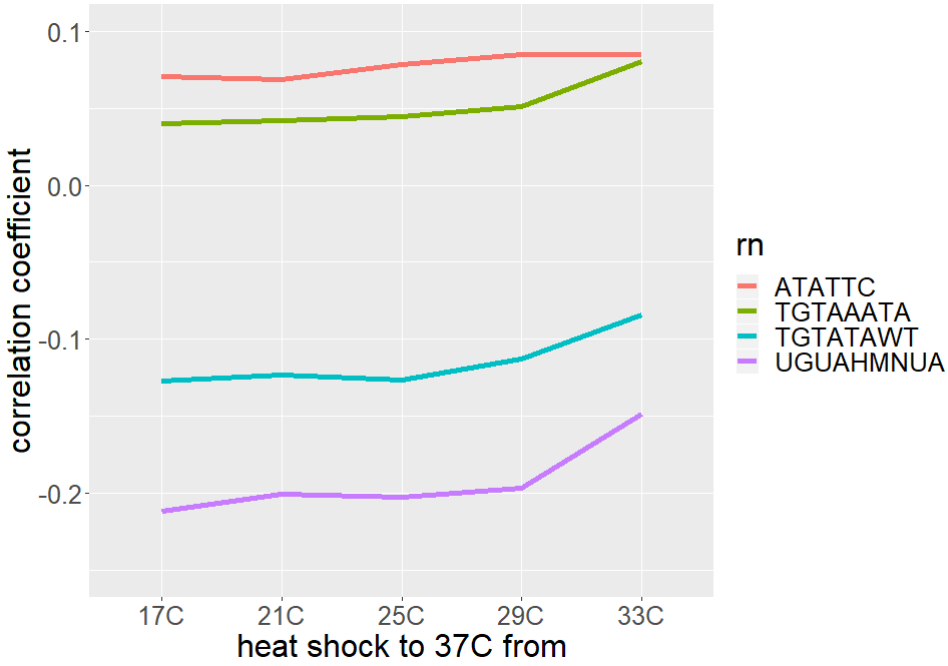


Figure 3-1. *The correlation coefficients between the frequency of the 4 heat-shock relevant motifs and the gene expression change after heat shock from various temperature to 37°C after 20min.The x-axis indicates the various degree of heat shock from 17°C, 21°C, 25°C, 29°C and 33°C to 37°C. The y-axis indicates the correlation coefficient between the frequency of the each of the four motifs and the log2 ratio of gene expression change. The curve shows how the correlation coefficient of the four motifs changes as the degree of environmental stress decreases.*

3.2 17 heat-shock-relevant motifs are selected from linear models

We temporarily removed the motifs with total frequency lower than 5 since we were not confident enough to tell whether a linear relationship exist with a small sample size. Then, we fitted the linear model on `heat shock from 25˚C to 37˚C after 15min` condition with the `lm()` function with default least squares method:

$$\min_{w} \sum (w * e)^2$$

17 heat-shock-relevant motifs were selected from the linear model with the following coefficient:

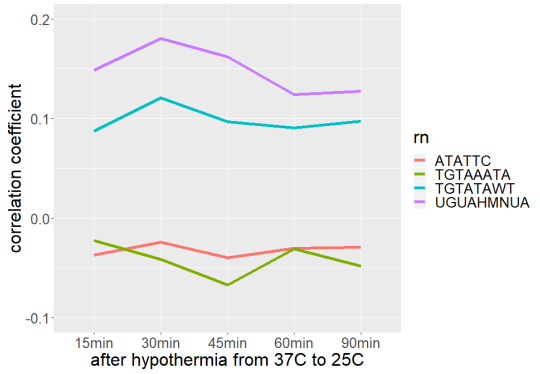
Motif	Estimate	Std.Error	t value	Pr(>|t|)
UGUAHMNUA	-1.041901588	0.05517823	-18.88247501	2.404205e-78
ATATTC	0.286249587	0.03720537	7.69377169	1.544898e-14
TGTATAWT	-0.394746133	0.06344118	-6.22223852	5.071069e-10
WUUGUAWUWU	-0.376281067	0.10137897	-3.71162838	2.068994e-04
TGTAAATA	0.247406481	0.06917155	3.57670844	3.493731e-04
UUUAAUGA	0.382709925	0.10796196	3.54485888	3.943794e-04
WUAUAUAW	0.099146169	0.03329943	2.97741365	2.912958e-03
CGCTATTG	1.023888345	0.37672006	2.71790237	6.579540e-03
GTAGTATTT	-0.941653721	0.37713600	-2.49685447	1.254388e-02
GTTTGTTG	-0.659629213	0.27719778	-2.37963382	1.734602e-02
WWTMGTATATTGTMA	-2.371028986	0.99751927	-2.37692550	1.747384e-02
AAAATAAAG	-0.401311025	0.17603931	-2.27966709	2.264563e-02
TATTATTAT	0.139838226	0.06206216	2.25319639	2.426548e-02
TCATGTAT	-0.268726894	0.12850339	-2.09120466	3.653156e-02
TATGTATTGT	-1.033652434	0.49833996	-2.07419135	3.808368e-02
ATGTAGTAG	-0.766822042	0.37684846	-2.03482863	4.189095e-02
CAGCAACT	0.784740042	0.39102954	2.00685615	4.478829e-02

Table 3-2. *Summary of linear model that predict how the gene expression change after heat shock from 25°C to 37°C after 15min. This table is sorted by* `Pr(>|t|)` *(also known as the p-value) which is the probability of achieving a |t| as large as or larger than the observed absolute t value if the null hypothesis (estimate = 0) was true. In short, it is the probability that the coefficient between factor and predict value is zero (Ronald L. Wasserstein et al., 2016).* `Estimate` *means the coefficient or weight gains by the responding features in this linear model.* `Std.Error` *is the standard error of the* `estimate` *value, which, can be used to calculate the Confidence interval. For example, the 95% confidence interval could be obtained from Estimate +- 1.96*Std.Error.* `t value` *is calculated from the estimates divided by their standard errors. To make the best use of this value, we need to look up the table of t distribution to learn the reject boundary. In this case, we could simply say the larger the magnitude of the t-value is, the less likely that the coefficient is 0.*

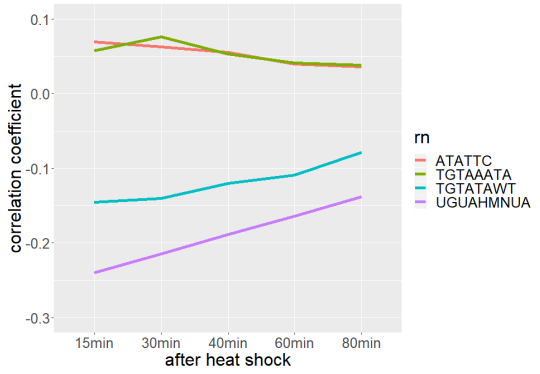
3.3 Identify potential relationships between environmental stress

In this part, we investigated whether there are potential relationships between the models represents different types of environmental stresses. After computing the correlation coefficient between the significant motifs and both the expression level change under hypothermia and heat shock, an opposite correlation could be found.

After hypothermia from 37C to 25C

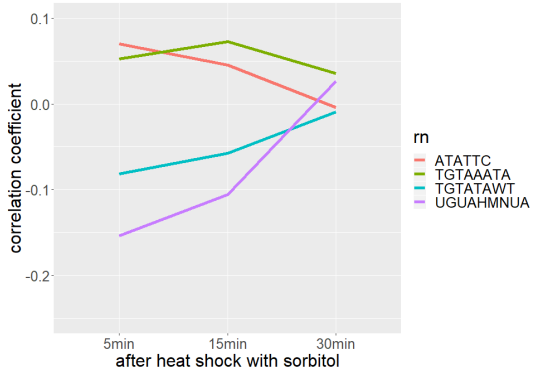


after heat shock from 25C to 37C

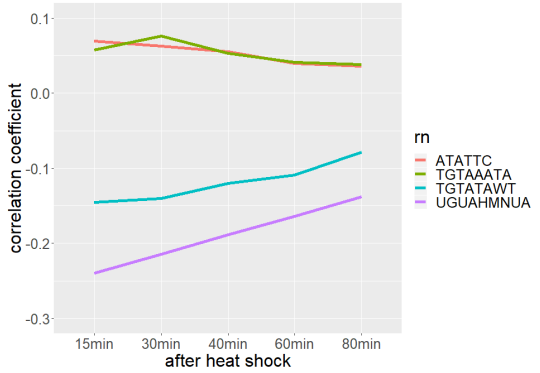


By computing the correlation coefficient between the significant motifs and both the expression level change under heat shock and heat shock with sorbitol. A similar correlation can be found.

heat shock with sorbitol



heat shock only



Chapter 4

Fit Group Lasso Model

To quantify how the cis-regulatory element affects the gene regulation under different environmental stresses, we fitted a series of linear models with group lasso where each model takes the frequencies of cis-regulatory elements as input and predicts how the gene expression level would change under certain environmental stress. The L2 penalty of group lasso can filter out most of the irrelevant factors by penalizing their coefficient towards zero. We imported `glmnet` packet and set the parameter `family` as `mgaussian` to fit multiple models at once and applied the elastic net penalty. The penalty on the coefficient vector for variable j is:

$$(1-\alpha)/2||\beta_j||_2^2 + \alpha||\beta_j||_2.$$

Lasso group is a special case of the elastic net. We set parameter α to 1 so that a group lasso penalty is used. As a result, the equation we want to minimize should be:

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{(p+1)\times K}} \frac{1}{2N}\sum_{i=1}^{N}||y_i-\beta_0-\beta^T x_i||_F^2 + \lambda\sum_{j=1}^{p}||\beta_j||_2$$

4.1 224 genes with expression profile missing rate larger than 10% is removed

The group lasso algorithm required a complete design matrix with no missing value. However, the expression level data from microarray usually contains lots of missing value due to insufficient resolution, image corruption, or simply due to dust or scratches on the microarray slide.

To begin with, we investigated the distribution of missing values from two aspects, (1) the proportion of the missing values from all genes in each environment and (2) the proportion of the missing value of specific genes under all environmental stresses. The results are summarized in Table 4-1 and Table 4-2. For each environmental stress, the proportion of missing values are less than 23%. The conditions with a relatively large missing proportion are mainly from heat shock and DTT exposure groups. On the other hand, 41 genes show an expression profile missing rate higher than 25%, which means for these genes, more than 44 of 173 expression profiles value are missing. Considering we planned to apply kNN imputation to fill the missing values where the more complete expression profile a gene has, the more accurate its neighbours could be found, we removed the 224 genes with missing values greater than 17 (10%).

number of environmental stresses	missing value (n)
0	0
152 (87.9%)	0 < n < 308 (5%)
7 (4.0%)	307 < n < 616 (10%)
14 (8.1%)	615 < n < 1385 (23%)

Figure 4-1. *missing value distribution of environmental stresses*

number of genes	missing value (n)
755 (12.3%)	0
4372 (71.1%)	0 < n < 9 (5%)
701 (11.4%)	8 < n < 18 (10%)
283 (4.6%)	17 < n < 44 (25%)
41 (0.7%)	43 < n < 99 (58%)

Figure 4-2. *missing value distribution of genes*

4.2 Fill the missing value of gene expression data with kNN imputation

In order to fill the rest of the missing data, we applied the kNN imputation recommended by Troyanskaya et al.[25] KNN imputation looks for the k most similar genes based on known expression profiles for each gene with missing values, then estimate the missing value from the expression profile of these neighbours. For example, if we want to estimate the missing value of gene A under `heat shock 5min`, we look for the k genes with similar expression profile as that of gene A under `heat shock 10min`, `heat shock 15min` etc. Then use their mean expression profile under `heat shock 5min` to fill the missing value.

We chose Euclidean distance, a metric shows a good and conservative performance in the examination of Troyanskaya et al., to measure the similarity between genes. Only the environmental conditions where the target gene has a non-NA values will be considered when searching for neighbours. Since we have removed genes with expression profile missing rate greater than 10%, at least 155 columns (90%) of the data can be used to find similar genes.

The hyperparameter k is selected by estimating the performance in mean square error. First, we randomly selected 1000 non-missing values from gene expression data as the first test group. Then changed the random seed and repeated nine times to obtain a total of ten sets of non-missing values. For each group, we removed the non-missing values from the matrix and performed kNN imputation together with the actual missing values. The returned estimated values were then compared with their actual values so that the mean square error could be calculated. When we repeated the estimating process with different k values, the ten sets of non-missing values were reused instead of selecting new groups of non-missing values for more reliable comparison. The results are shown in Figure 4-3.

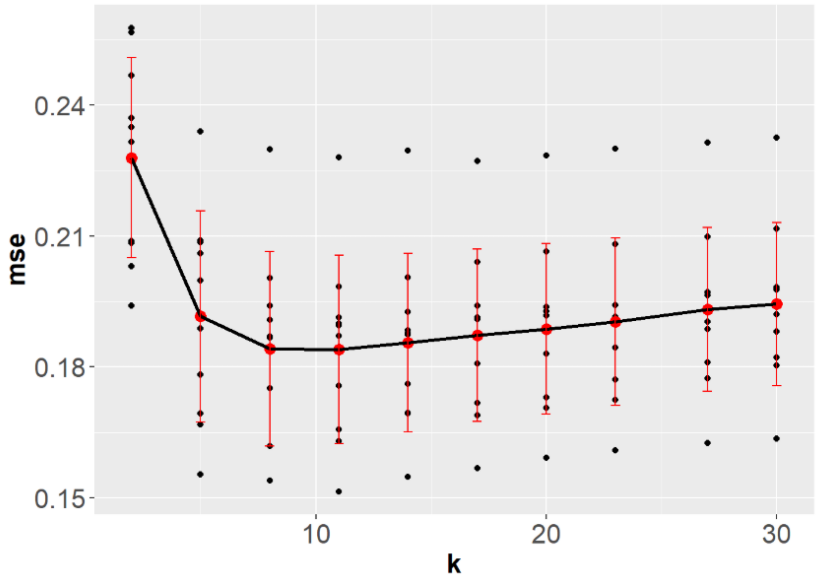


Figure 4-3. *The estimated mean square error of kNN imputation with a series value of k. For each value of k, the performance estimation have been carried out for ten times where each black dot represents one estimation and the red dots represent the mean value of mean square error of each value of k. The error bars indicate 1 standard error of the mean.*

According to Figure 4-3, the optimal value for k should sit between 8 and 11. The poorness of performance under a lower k value is probably caused by the overemphasis of a few dominant expression patterns. On the other hand, the deterioration in performance with larger values of k could not only cause by the distance increase between target gene and its neighbours. There is usually noise presented in the data from the microarray. As k increase, the contribution of noise overwhelms the contribution of the signal and thus, reduces the performance.

4.3 Compare the estimate performance of kNN imputation with baseline

In order to evaluate the performance of kNN imputation, we created two baselines, 1) filling the missing values with the mean of all non-missing values under the same conditions and 2) directly setting the missing values to 0. The former is a generally used performance-priority baseline. The other one takes the assumptions of the current situation in account: When the log2 ratio that represents the relative transcript level is 0, it means the gene expression level does not change.

Then we compared the estimated performance of kNN imputation with that of the two baselines (Figure 4-4). We applied the same method as the kNN imputation evaluation to evaluate the performance of the baseline. The ten sets of random values used to evaluate the performance of KNN were reused for a more general comparison. Because the kNN imputation provides not only much better but also more conservative performance, we filled the missing values in the design matrix with the value estimated by kNN imputation (k = 11).

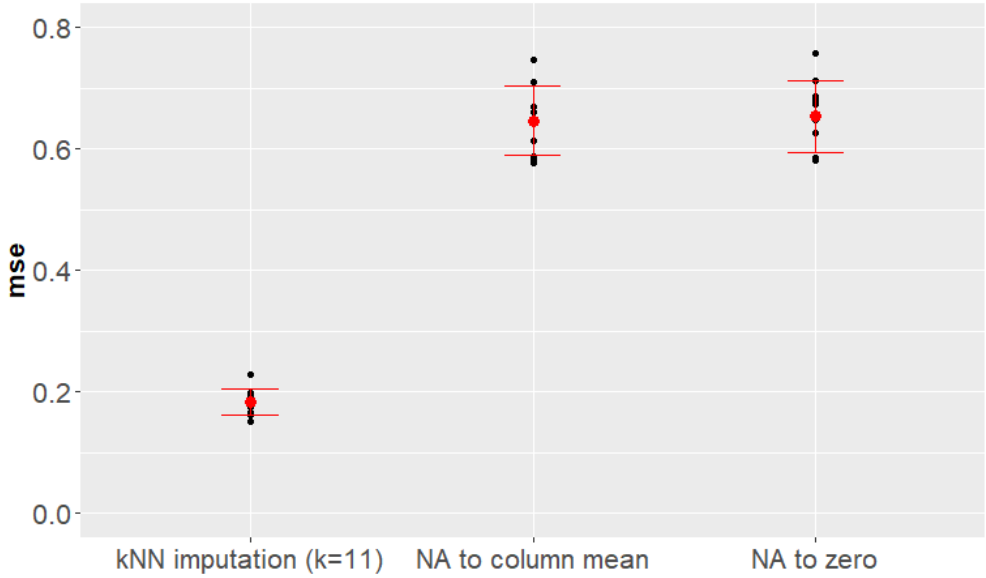


Figure 4-4. *The estimate mean square error of kNN imputation and baselines*

4.4 Fitting group lasso model with hyperparameter found by cross-validation

λ is the hyperparameter that adjusts the effect of the penalty. A small λ could indirectly lead to the overfitting of the model while a large λ may not only lead to a deterioration in the performance but could also filter out many important factors. As a general approach, we applied the 10-fold cross-validation provided by the `glmnet` package [26] to select the lambda value that can give both good performances and retain most of the factors. Two lambda value were returned from the function `cv.glmnet`, named `lambda.min` and `lambda.1se`. `lambda.min` is the value of lambda that gives minimum cvm(mean cross-validated error) and `lambda.1se` is the largest value of lambda such that error is within 1 standard error of the minimum.

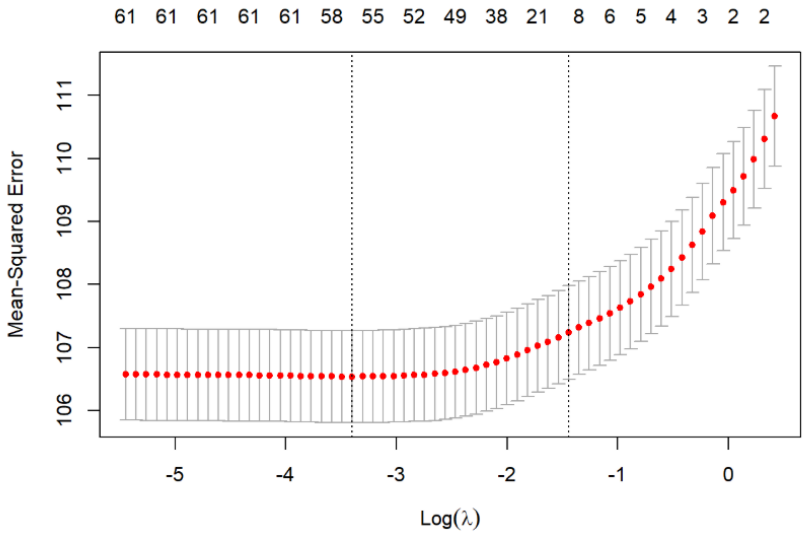


Figure 4-5. *Mean-squared Error of cross validation under different value of λ. The two vertical line correspond to* `lambda.min` *and* `lambda.1se` *respectly. The confidence intervals represent error estimates for the loss metric (red dots). The vertical dashed lines show the locations of λmin and λ1se. The numbers on the top are the estimated number of nonzero coefficient. The error is accumulated, and the average error and standard deviation over the folds are computed.*

According to figure 4-6, training model with λ = `lambda.min` results in a model with the highest expected performance while training model with λ = `lambda.1se` results in a more conservative model with much fewer non-zero regression coefficients.

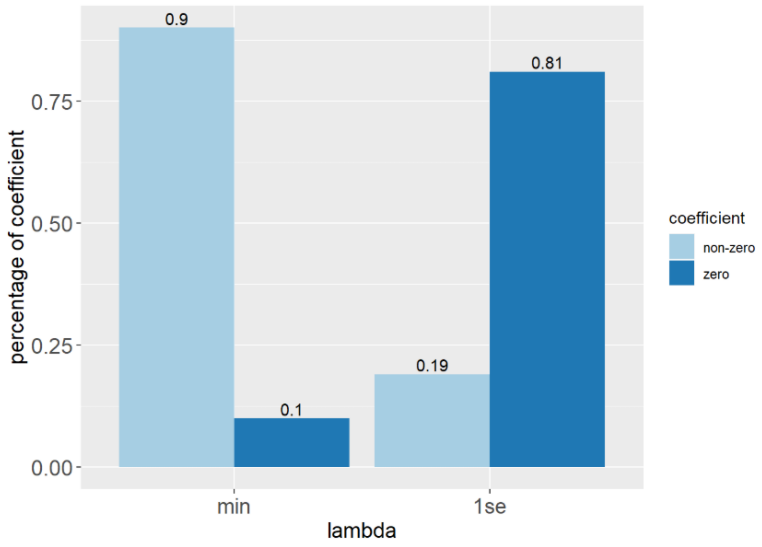


Figure 4-6. *comparison of the number of non-zero coefficient between models training with two lambda value*

Although the `lambda.1se` is chosen in most of the relevant studies, we prefered `lambda.min` for our model since our target is to quantify the regulatory effect of motifs and to figure out motifs with different regulatory effect under different environmental stresses. Moreover, the motifs we applied in our model were imported from previous studies and their effect have been validated with various methods. Thus, It is not suitable to remove them in large quantities.

Chapter 5

Analyze the regression coefficients

There are 173 models fitted with group lasso where each one explains how the motifs relate to the expression level change under different environmental stresses. Their coefficients can be used to estimate the regulatory effect of cis-regulatory elements. For example, if an element is given a relatively large coefficient in a model which predicts the change of expression level under certain environmental stress, we can say the element play a relatively important role in the regulation under that environmental stress. Furthermore, even the bias of the model provides clues about how important the roles played by these motifs under the overall gene expression regulation. Since the bias can be understood as the expected gene expression level change of a gene with no cis-regulatory element we considered on its 3'UTR sites, the bias indicates the expression level change that cannot be explained by the model.

5.1 Group conditions by the type of environmental stress

Firstly, we selected the models we interested in and divide them into 10 groups. The environmental stresses included in each group are listed in appendix:

1. heatshock from 29°C to 33°C
2. heatshock from 29°C to 33°C with sorbitol
3. heatshock from 25°C to 37°C
4. hypothermia from 37°C to 25°C
5. sorbitol treatment
6. nitrogen depletion
7. diamide treatment
8. dithiothrietol exposure
9. menadione exposure
10. hydrogen peroxide treatment

Above all types environmental stresses, we are most interested in `heatshock from 29°C to 33°C`, `sorbitol treatment` and `heatshock from 29°C to 33°C with sorbitol` due to their relationship. Its worth to mention that `sorbitol` is a sugar that acts as an osmoprotectant which can be used to separate effects of heat internal to the cell with effects on osmotic pressure.

5.2 The regulatory effect of motifs peak at different time under different type of environmental stress

By extracting the regression coefficients given to motifs and sorting them with the previous grouping, we found that the regression coefficients of most motifs show a tendency to rise first, peak in a certain period and then fall. It describes the process of the regulation effect which gradually increases and returns to a stable state after the environmental stress. The time when the regulation effect reaches its peak varies with the type of environmental stress. For example, undereat shock-related environmental stress, the regression coefficients of most motifs peaked in about 15 minutes. while under diamide treatment, the regression coefficient of motifs generally reaches its peak after 30 minutes. It is worth mentioning that the degree of environmental stress has no obvious effect on the peaking period. As shown in Figure 5-1 and Figure 5-3, for most of the motifs, both their regression coefficients in `heat shock from 25°C to 37° C` and `heat shock from 29°C to 33° C` peaked at around 15 minutes. Moreover, even under the same environmental conditions, not all the regression coefficients peak at the same period. This may because of the complex regulatory mechanism within the cell. While the cell adjusts its gene expression level, the number of regulatory elements (as products of the gene expression) such as RNA binding proteins that participate in regulation also changes, and lead to a change in the regulatory effect of motifs change as well.

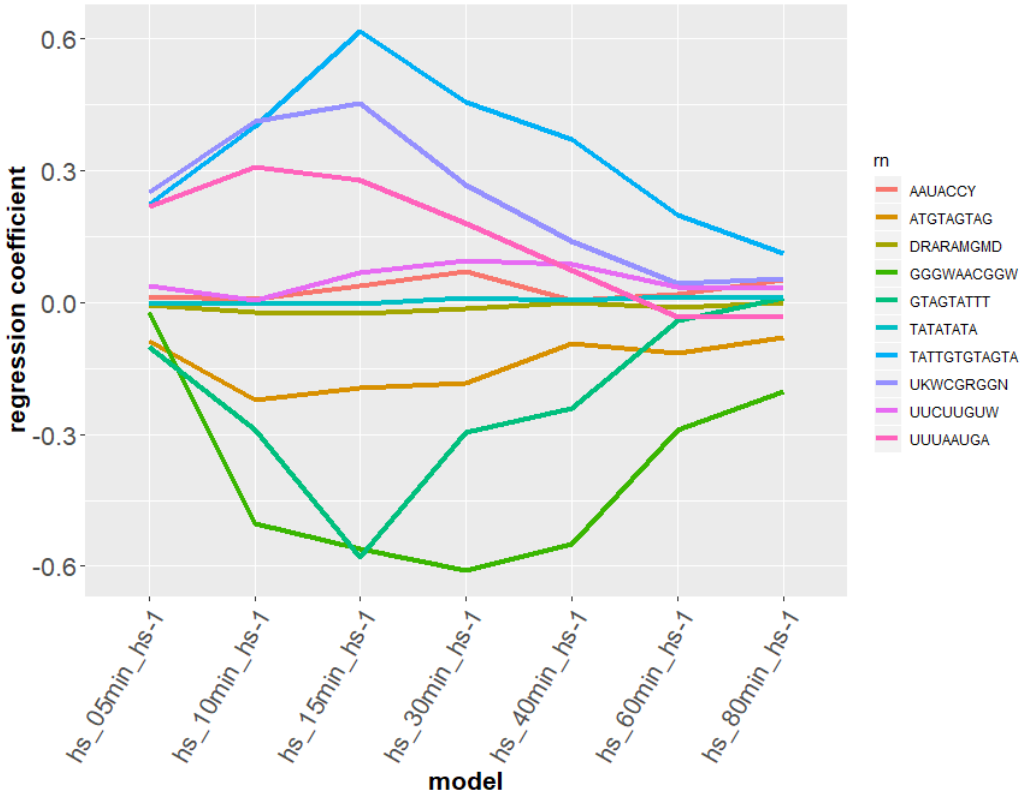


Figure 5-1. *The regression coefficient of 10 motifs from models predicted how the gene expression level change after heat shock from 25°C to 37°C. For most of the motifs, their coefficients increase with a short delay and peak at 15mins. A possible explanation of the short-time delay is that the majority of cis-regulatory elements on 3'UTR regulate the expression level by adjusting the stability of mRNA and it takes some time to reflect the regulatory effect on the transcript level.*

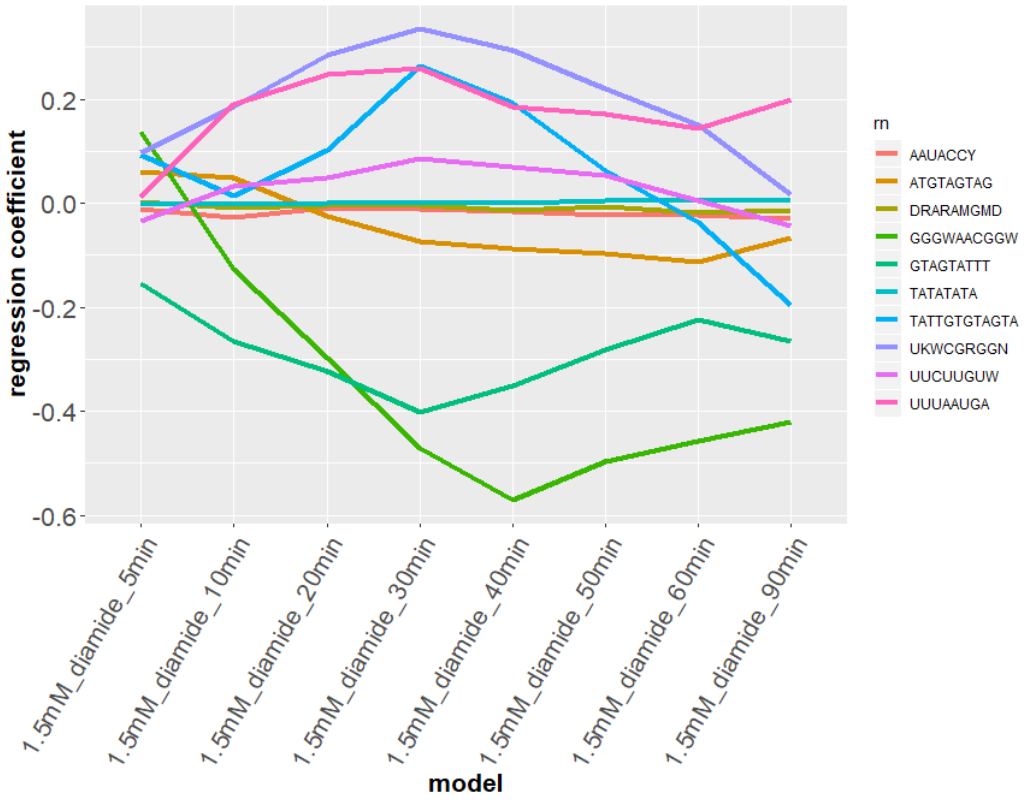


Figure 5-2. *The regression coefficient of 10 motifs from models predicted how the gene expression level change after diamide treatment. Most regression coefficient peak after 30mins.*

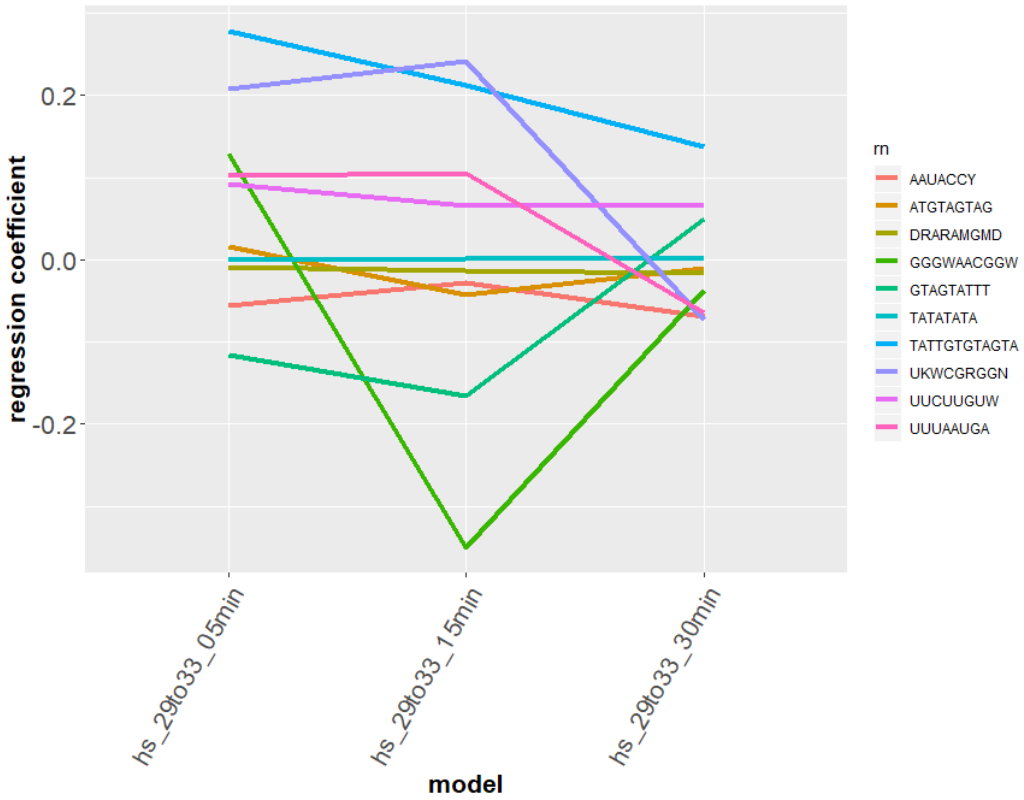


Figure 5-3. *The regression coefficient of 10 motifs from models predicted how the gene expression level change after heat shock from 29°C to 33°C.*

Therefore, we estimated the period when the gene expression effect is biggest under each type of environmental stress by comparing the variance of the log2 ratio of gene expression level. Time slice with relatively high variance is considered as the period when the gene expression effect is biggest.

type of environmental stress	period when gene expression effect is biggest (min)
heatshock	15 ~ 20
hypothermia	15, 60
sorbitol treatment	30 ~ 45
nitrogen depletion	2880 ~ 7200
diamide treatment	30 ~ 40
dithiothrietol exposure	20 ~ 60
menadione exposure	30 ~ 40, 80~120
hydrogen peroxide treatment	20 ~ 30

Table 5-4. *the period when the regulatory effect of most motifs are significant*

5.3 Motifs perform regulatory effect differently on heat shock under sorbitol condition

We have been looking for a reliable way to compare how different motifs contribute to transcriptional regulation under different types of environmental stresses. As an attempt, we carefully selected the condition with overlapping time slice in `heat shock` and `heat shock with sorbitol` for comparison:

- heat shock
  - hs_29to33_05min
  - hs_29to33_15min
  - hs_29to33_30min
- heatshock with sorbitol
  - 29C(1M_sorbitol)~33C(1M_sorbitol)_05min
  - 29C(1M_sorbitol)~33C(1M_sorbitol)_15min
  - 29C(1M_sorbitol)~33C(1M_sorbitol)_30min

The regression coefficients of 61 motifs from 6 models, that is, a total of 6 * 61 coefficients were summarized and compared. To represent the contribution of motif m under stress group t, we have considered four types of metrics `mean`, `mean of absolute value`, `L2 norm` and `mean L2 norm`:

$$C_t : \text{set of conditions from the type of environmental stress } t$$
$$\beta_{mc} : \text{regression coefficient of motif m under condition } c$$

$$mean = \frac{\sum_{c \in C_t} \beta_{mc}}{|C_t|}$$

$$mean\ of\ absolute\ value = \frac{\sum_{c \in C_t} |\beta_{mc}|}{|C_t|}$$

$$||\beta||_2 = \sqrt{\sum_{c \in C_t} \beta_{mc}^2}$$

$$mean\ ||\beta||_2 = \sqrt{\frac{\sum_{c \in C_t} \beta_{mc}^2}{|C_t|}}$$

Mean and median are widely used parameters for describing distribution. However, we are more interesting in the strength of the regulatory effect of motifs rather than whether it represses or promotes gene expression. Assuming there is a motif has a strong repression effect on gene expression at the first 10 mins and then promotes it slightly, we tend to classify it as a motif with significant regulatory effect. In this case, the mean of its coefficient is likely to be close to 0, so that the effect of the motif is underestimated.

Therefore, we considered applying the `mean of absolute value` and `L2 norm` as a metric of the strength of the regulatory effect. The former is more meaningful as it indicates the average contribution of motif to the gene expression regulation. While `L2 norm` is more focusing on the motifs with strong regulatory effect at a certain period. Instead of applying both `mean of absolute value` and `L2 norm` to indicate the contribution of motif, we introduced `mean L2 norm`, which gives a relatively meaningful value that larger than mean and smaller than the maximum coefficient. For motifs with same `mean absolute value`, those with stronger regulatory effects in a relatively short period have a larger `mean L2 norm`.

The `mean L2 norm` of regression coefficients from both group are compared and order with transcript frequency as a reference of reliability:

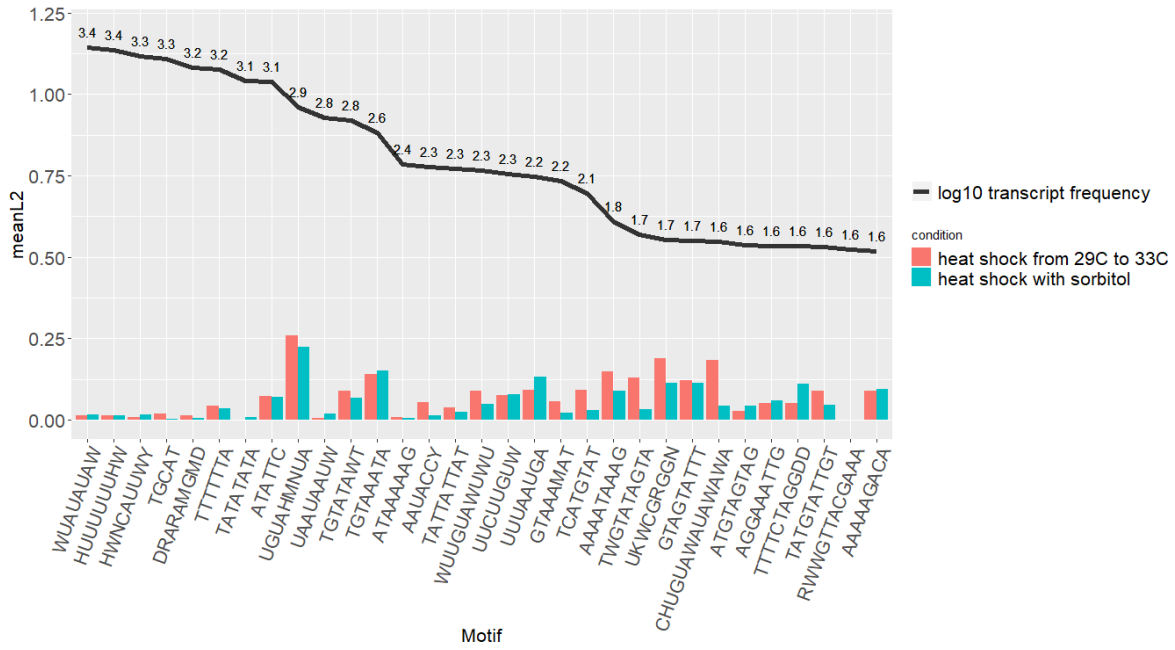


Figure 5-5 (A). *transcript frequency and mean L2 norm of each motif*

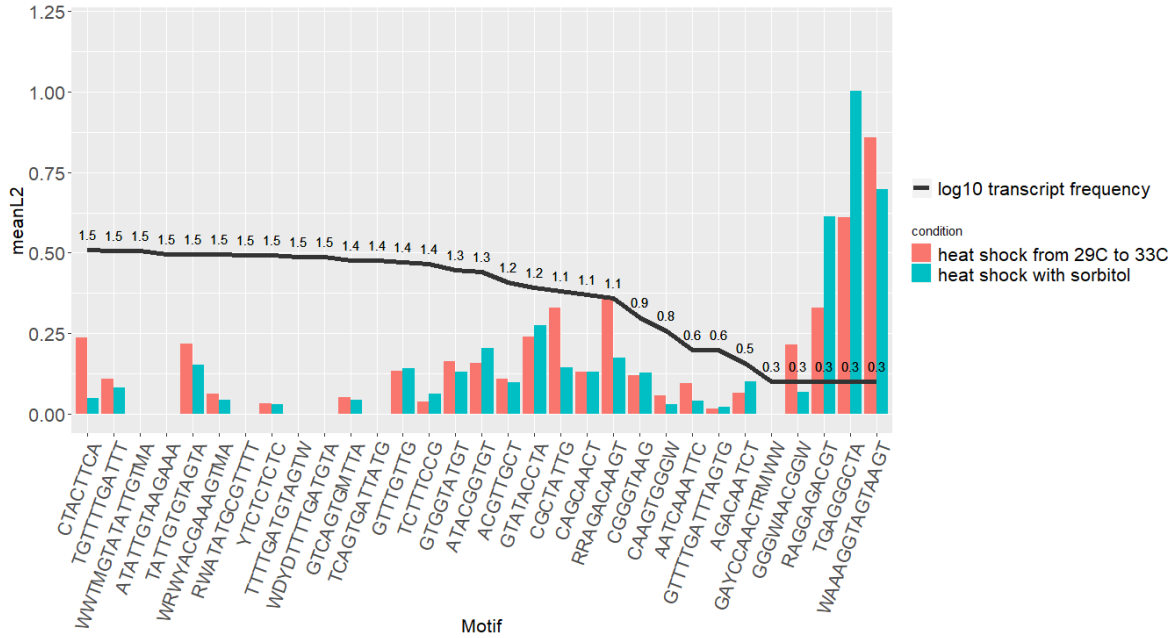


Figure 5-5 (B). *transcript frequency and mean L2 norm of each motif*

Figure 5-5 shows the comparison between the `mean L2 norm` of regression coefficients from condition groups `heat shock` and `heat shock with sorbitol`. The black line indicates the transcript frequency corresponding to each motif. It should be noted that the transcript frequency here refers to the number of genes that have that specific motif in 3'UTR, not the totaloccurrences of that motif in all 3'UTR. In this case, we believed the prediction on motif with

low transcript, such as `GGGWAACGGW`, `RAGGAGACGT`, `TGAGGGCTA`, `WAAAGGTAGTAAGT`, is unrealiable. A few motifs show a dramatically different regulatory effect when sorbitol exists. The `mean L2 norm` of `AAUACCY`,`GTAAAMAT`, `TCATGTAT`, `CHUGUAWAUAWAWA` and `CTACTTCA` under `heat shock with sorbitol` is much smaller than that under `heat shock`.

We believed the differences in the regulatory effect of motifs under different environmental stresses are common. However, it could be hard to apply the `mean L2 norm` comparison between each group of environmental stresses since the overlapping of time slice between each condition group is limited and may not cover the period when motifs show their maximum regulatory effect. Moreover, the value of `mean L2 norm` depends on how strong the environmental stress is. Stronger environmental stress usually leads to a more significant change in gene expression level and results in larger regression coefficients as well as the `mean L2 norm`. Thus, we kept looking for an alternative approach.

5.4 6 motifs with relatively high expalined varaince are selected

As we have mentioned previously, the overall regulatory effect in a cell also depends on how strong the environmental stress is. For example, the change in expression level under heat shock from 25 degrees to 37 degrees should be much larger than that under heat shock from 29 degrees to 33 degrees. Therefore, to quantify the regulatory effect of these motifs under the overall regulatory process, we calculated the percentage variance explained by each of them under different types of environmental stresses.

$$C_t : \text{set of conditions from the type of environmental stress } t$$
$$\beta_{mc} : \text{regression coefficient of motif m under condition c}$$

$$e_{mg} = \begin{cases} 1, & \text{at least one motif } m \text{ is on gene } g \\ 0, & otherwise \end{cases}$$

$$\begin{aligned} \text{Variance predicted by motif} &= \sum_{c \in C_t} \sum_{g} \beta_{mc}^2 \times e_{mg} \\ &= \sum_{c \in C_t} \beta_{mc}^2 \times transcript\ frequence \\ &= ||\beta_{mc}||_2^2 \times transcript\ frequence \end{aligned}$$

In order to select motifs with high explained variance, we calculated the total variance explained by each motif under all condition (Figure 5-7). As a result, 6 reliable significant motifs `UGUAHMNUA`, `ATATTC`, `TGTATAWT`, `TGTAAATA`, `ATACGGTGT`, `WUUGUAWUWU` are selected.

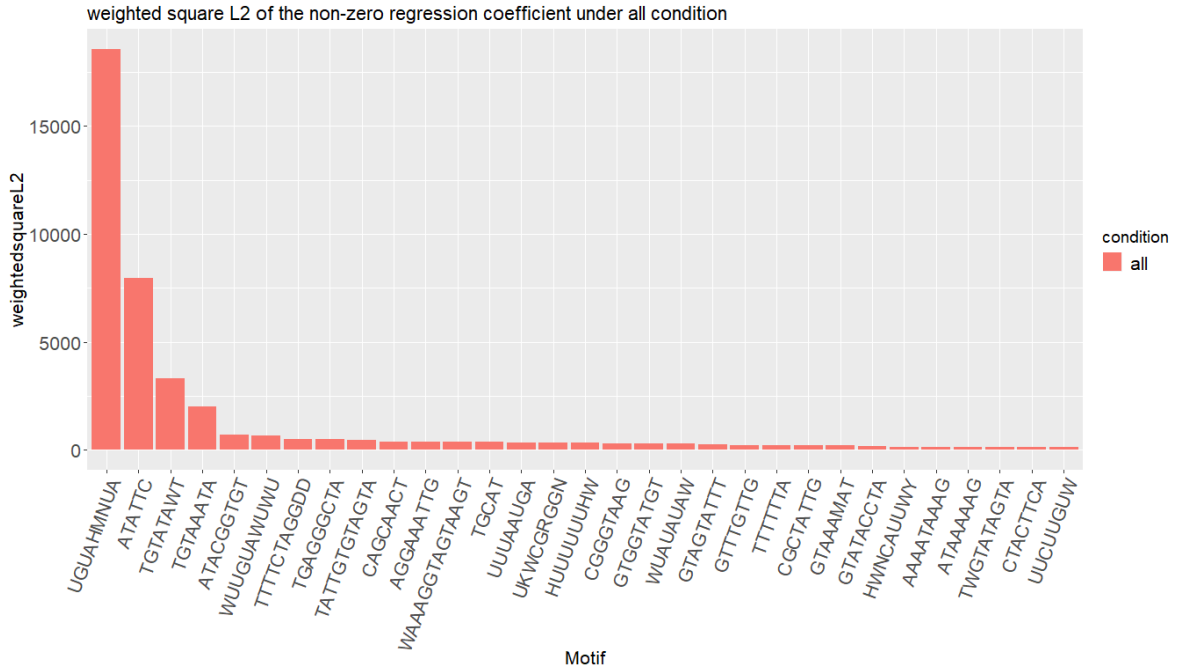


Figure 5-7. *total variance explained by each motifs under all condition*

5.5 Compare the percentage explained variance of 6 significant motifs under each group of environmental stress

We focused our investigation on the six selected significant motifs and plotted bar charts to compare the proportion of variance explained by each of them across each group of environmental stresses. The bar charts have provided lots of useful information. In the first place, we expected the proportion of variance explained by motif under `heat shock from 25˚C to 37˚C` and `heat shock from 29˚C to 33˚C` be similar since they belong to the same type of environmental stress. However, according to the plot, the proportion of variance explained by motifs under stronger environmental stresses is greater. Furthermore, we have calculated the variance explained by motifs under both `heat shock from 29˚C to 33˚C`, `sorbitol treatment` and the compound of them - `heat shock from 29˚C to 33˚C with sorbitol`. As long as the motif responds to each type of the environmental stress respectively, we could expect the proportion of variance explained by motif under the compound environmental stress `heat shock + sorbitol` sits between that under `heat shock` and under `sorbitol treatment`. While the proportion of variance explained by motif `TGTAAATA` under `heat shock + sorbitol` is actually much greater than that under `heat shock` or `sorbitol treatment`.

$$\text{proportion of variance explain by motif } = \frac{\text{variance explain by motifs}}{\text{total variance}}$$

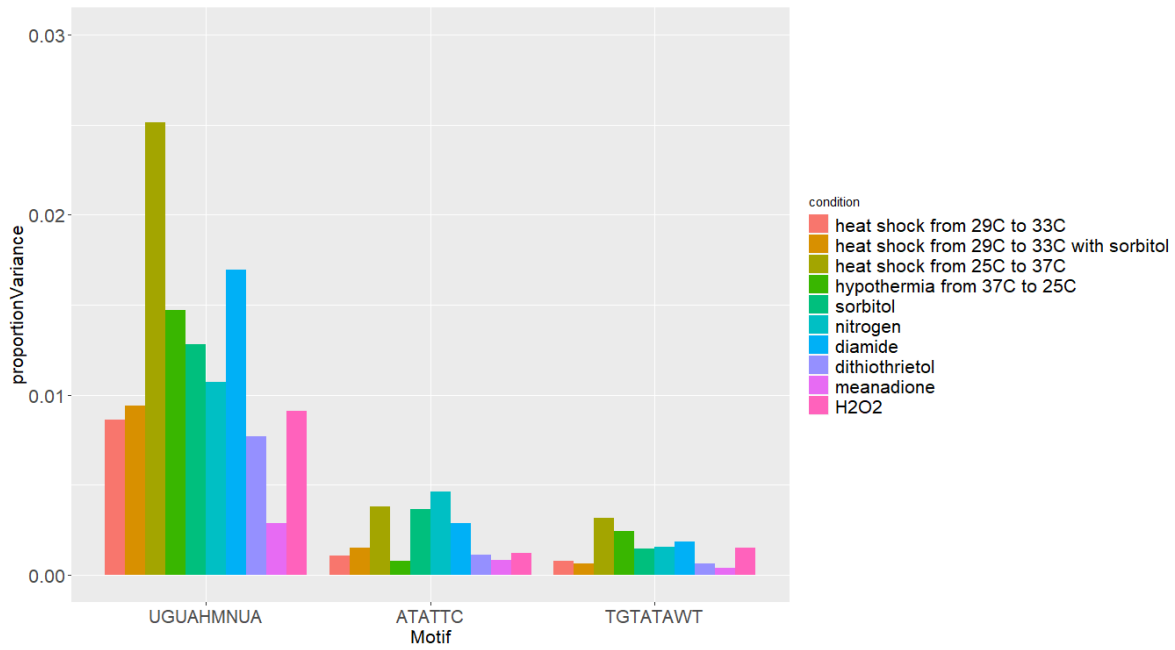


Figure 5-8 (A). *proportion of motifs explained variance*

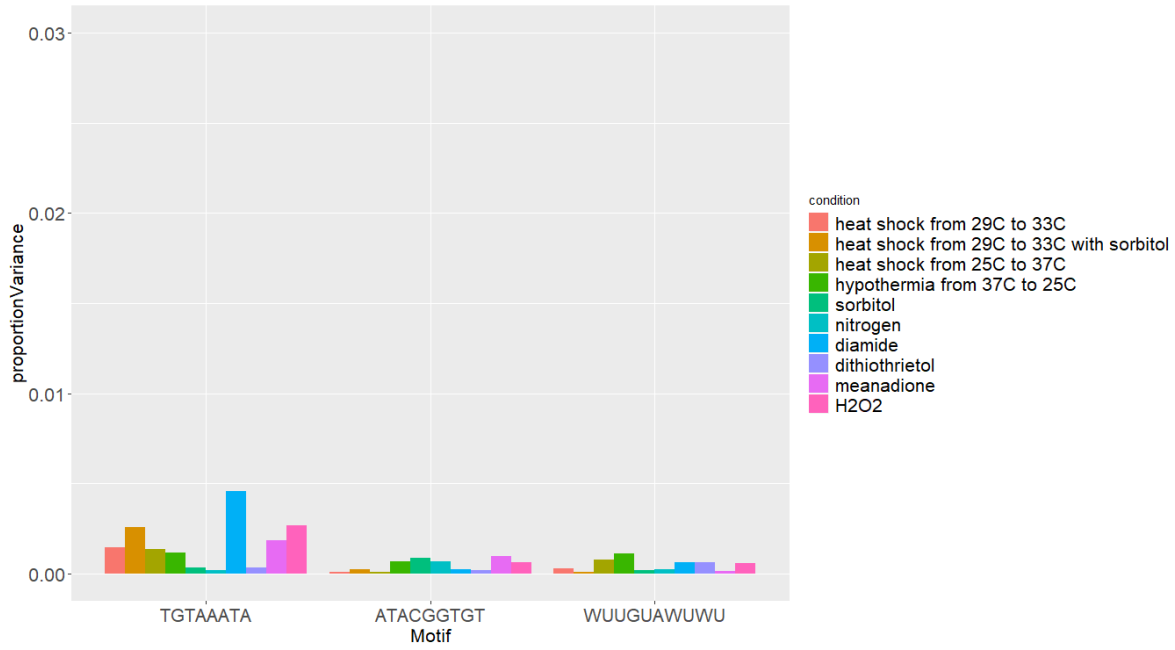


Figure 5-8 (A). *proportion of motifs explained variance*

5.6 Evaluate the reliability of the regression coefficient of the 6 significant motifs

Because the model is already fitted greedily with group lasso, the traditional method like using chi-sequared test to compare the drop in residual sum of squares to X^2 distribution is no longer appropriate: adaptivity makes the drop in residual sum of squares stochastically much larger than X^2 under the null hypothesis.

As an alternative approach to evaluating the reliability of the regression coefficient of significant motifs, we fitted a linear model with selected motifs and investigate their regression coefficient and standard error. According to previous analysis of the regulatory effect at different times, we applied the most representative gene expression levels within the period when the regulatory effect of motifs is strongest. We fitted 10 models in total and plot their coefficient and standard error. The error bars represent the range of 2 standard errors. (Figure 5-9)

Most of the regression coefficients of `ATACGGTGT` and `WUUGUAWUWU` are considered not significant enough since their 95% confidence intervals sit across zero.

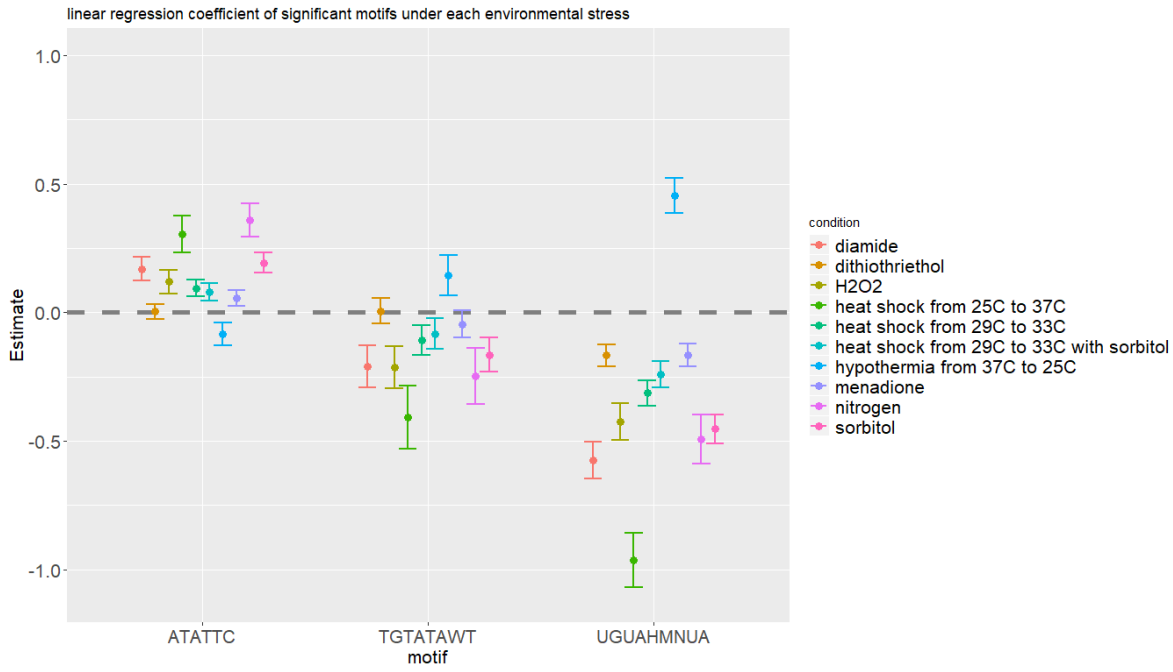


Figure 5-9 (A). *regression coefficient of significant motifs from linear model*

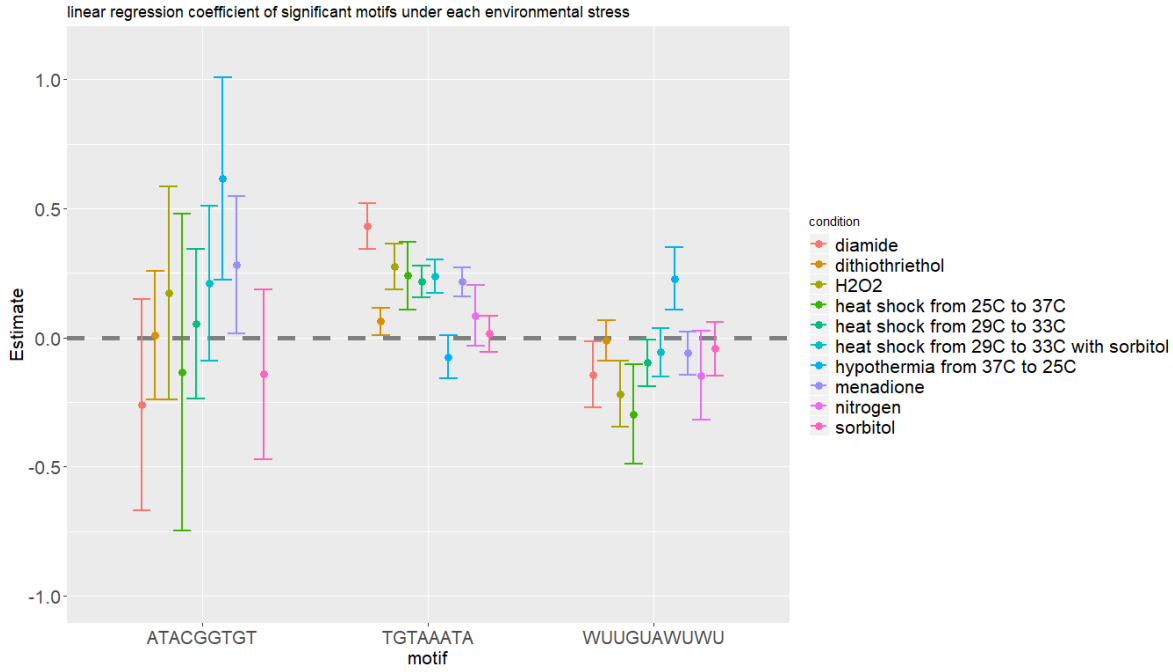


Figure 5-9 (B). *regression coefficient of significant motifs from linear model*

Chapter 6

Conclusion and Discussion

In this section, we summarized the previous finding, propose motifs worth experimentally verifying and discuss the benefit and problem with our approaches.

6.1 The stronger the environmental stress, the greater regulatory effect of motif have

By comparing the `heat shock from 29C to 33C` and the `heat shock from 25C to 37C` in the variance explained by significant motif, it is obvious that the six relatively significant motifs can explain more variation when the temperature difference increases, which is not like what we expected: since both groups of environmental stress are heat shock, we expected the percentage of variance explained by motifs to be similar.

One possible explanation is that when environmental changes intensify, cells will rely more on post-transcriptional regulations for a stronger gene expression effect. The cells will be eager to adjust the amount of mRNA within. However, it is not enough to simply speed up or low down the mRNA synthesis rate. It would be necessary to adjust the mRNA's decay half-life as well.

Another possible explanation does not conflict with the first one but could be a reason that leads to an overestimation of the first interpretation: the noise during the measurement makes us underestimate the regulatory effect of each motif under milder environmental stress. Suppose there is a fixed noise for all measurements of the gene expression level. Under milder environmental stress, because the change in gene expression level is smaller, the noise will be the cause of the majority variance. Since these variations cannot be explained by motifs it leads to more significant differences in the predicted regulatory effects of the motifs under milder environmental stress and stronger environmental stress.

6.2 Motif UGUAHMNUA occupies a relatively important position in the post transcriptional regulation

According to the comparison data of variance predicted by motif and significant rank, it is not difficult to see that above all the reliable motifs, the motif UGUHMNUAA contributes a much larger regulatory effect than others. It is still predicted to have a greater effect on gene expression in a variety of environments with a relatively high transcript frequency even it has a relatively high transcript frequency. Even so, `UGUAHMNUA` still has weak regulation ability under menadione exposure. In fact, similar feature were observed in many other motifs. We suggest two possible explanations for this:

1. Cells do not need an extensive gene regulation to respond to menadione exposure

2. Cells do not rely much on Post-transcriptional regulation to respond to menadione-related environmental stress

In a study on RNA recognition elements, Daniel et al. predict `UGUAHMNUA` as a binding site of the regulator factor Puf4 by analyzing the conserved sequences and conservation rates [32]. This, to some extent, confirms our model's ability to predict the regulatory effect of motifs and allows us to extend the prediction of the motif function to the protein associated with it.

On the other hand, `TGTAAATA`, which has a similar heading but different ending comparing with `UGUAHMNUA`, is said to be another transcription factor from the PUF family [33]. It is surprising that similar motifs, which are recognised by similar and related proteins, have such different effects.

6.3 Motif ATATTC shows a strong regulary effect on head shock relavant stress while much less effect on hyporthemia

Based on the analysis of the linear model and the correlation coefficient between motif and the change of gene expression, most motifs regulate gene expression under heat shock and hypothermia in the opposite way. However, indicated by the comparison of variance explained by motifs, `ATATTC` seems to contribute a much stronger regulate effect under heat shock rather than hyperthermia. Although some other motifs show similar behaviour, their difference of variance explained between heat shock and hypothermia is not as large as `ATATTC` does.

6.4 Motif TGTAAATA contributes a greater regulatory effect under heat shock with sorbitol than it does under heat shock without sorbitol even it has little regulatory effect under sorbitol treatment

By comparing variation explained by motif and motif significant rank, we can conclude that motif `TGTAAATA` has a robust regulatory contribution in heat shock and almost no ability to regulate gene expression under sorbitol treatment. However, when cells are exposed to both sorbitol and heat shock at the same time, the regulatory effect of motif `TGTAAATA` was enhanced. It could be interesting to investigate how motif `TGTAAATA` responds to heat shock and sorbitol exposure individually and together.

6.5 Group lasso makes it possible to summarize the regulatory effect of each motif under different environmental stress

The comparison of the regulatory effect of one motif across environmental stresses is difficult because the regulatory effect of one motif could be activated at different time periods under different types of environmental stress. With the multidimensional regression coefficients from group lasso, we are able to obtain an overview of the regulatory effect of each motif by computing the sum, L2 norm, mean L2 norm or variance explained by factors, etc.

6.6 Grouping by motifs may filter out motifs that work only in a few environmental stress

All of the motifs we selected from previous studies are considered to have a gene regulatory effect and contribute more or less in response to various environmental stresses. However, there may be specific motifs involved in gene regulation only regulate gene expression under some specific environmental stresses. According to the penalty formula of the group lasso, the regression coefficients in the same group will be penalized to zero simultaneously. In our study, we grouped the regression coefficients by motifs. Thus, the regression coefficients of the motifs only respond to a few types of environmental stresses are likely to be penalized to 0. In other words, it is difficult for us to find the motifs that focus on specific environmental stresses.

On the other hand, the imbalance of gene expression data may lead the model to prefer certain types of environmental conditions. Of the 173 sets of gene expression data we cited from the data from Gasch, 43 are related to temperature change. While only 9 of them were associated with menadione exposure. This may cause more temperature-dependent motifs being selected when fitting model.

6.7 Grouping by condition could be another approach

Although we could no longer filter out relevant motifs by group lasso when grouped by environmental condition, it can definitely help us avoid the problems mentioned earlier. The regression coefficients of the motifs that only responds to a few types of environmental stress won't be penalized together. Instead, the regression coefficient of the model that predicts gene expression under environmental stress which is hard to be explained by selected motifs is likely to be penalized to zero.

However, this raises another problem that the model may underestimate the effect of the motifs that show their regulatory effect at a different period of most other motifs. Referencing the previous finding - not all the motifs maximize its regulatory effect at the same time. For example, under heat shock, the regulatory effect of `TGTAAATA` maximizes at 30 min while that of most of the other motifs maximizes at 15min. When we group by environmental stimulus type, the regression coefficient representing 30 minutes is more likely to be penalized stronger than that representing 15 minutes and results in the underestimation on the regulatory effect of `TGTAAATA`.

6.8 Bayesian hierarchical approach

We believe the Bayesian hierarchical approach could be an excellent alternative approach for this project, although it is much more expensive than the group lasso. It has unparalleled advantages in the significance test and comparison of the similarity between models.

Applying the Bayesian approach means that it is not necessary to carry out the significance test on the regression coefficient. Because in the Bayesian model, the coefficients are chosen through the posterior probability, which is derived from the data distribution and prior probability of coefficient. The posterior probability can be estimated with Bayes' theorem:

$$p(w) : \text{prior probability of coefficient}$$
$$p(w|D) : \text{posterior probability of coefficient given data}$$

$$p(w|D) \propto p(D|w)p(w)$$

In a two-level hierarchical model, the posterior probability should be:

$$p(w, \theta|D) \propto p(D|w, \theta)p(w, \theta) = p(D|\theta)p(\theta|w)p(w)$$

Once the posterior probability is obtained, we can not only easily evaluate the reliability of the parameter, but also get useful information like the reliable range and trend, etc. of the coefficients.

Furthermore, the hierarchical model allows us to set similarity between models representing different environmental stresses and analyze the difference and similarity between models individually. The priority probability p(w) as a hyperparameter also ensures the regularization of model. A reasonable priority probability, like a penalty, can prevent overfitting.

However, it could be difficult to implement the Bayesian hierarchical model on our data. According to the background of our research, the posterior probability distribution p(w|D) is very likely to be far from Gaussian. Therefore, it is unsuitable for applying methods like Laplace approximation which assume the posterior probability follow Gaussian.

Appendix

IUPAC nucleotide code

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base

Motifs list

ATATTC
AAAAAGACA
TGCAT
CAAGTGGGW
TGTAAATA
CTACTTCA
TTTTTTA
TGAGGGCTA
CHUGUAWAUAWAWA
YTCTCTCTC

UGUAHMNUA
AGACAATCT
WUUGUAWUWU
TGTATAWT
HUUUUUUHW
TATGTATTGT
UAAUAAUW
GTGGTATGT
AKUCAUUCCUU
TGGGTGGGTA
WUAUAUAW
CGCTATTG
HWNCAUUWY
TCATGTAT
DRARAMGMD
TAAAAAGTAAAC
GGGWAACGGW
CAGCAACT
AAUACCY
AATCAAATTC
UKWCGRGGN
ATACGGTGT
UUUAAUGA
ACGTTGCT
UUCUUGUW
TCTTTCCG
TATATATA
TTTTCTAGGDD
TATTGTGTAGTA
GTTTGTTG
ATGTAGTAG
RRAGACAAGT
GTAGTATTT
WWTMGTATATTGTMA
GTAAAMAT
WRWYACGAAAGTMA
TTTTGATGTAGTW
KCTTGGAGGRRR
TGATTTAGTGTTTGT
YAWTKAGGGCTATK
GTATACCTA
TWGTATAGTA
RAGGAGACGT
AAAATAAAG
GTTTTGATTTAGTG
GTCAGTGMTTA
TATTATTAT
RWWGTTACGAAA
TGTTTTGATTT
WDYDTTTTGATGTA
TCAGTGATTATG
TTGAACATCCG

AGGAAATTG
RWATATGCGTTTT
ATATTGTAAGAAA
GAYCCAACTRMWW
CGGGTAAG
WAAAGGTAGTAAGT
ATAAAAAG

Environmental condition groups

1. heatshock from 29°C to 33°C
    - hs_29to33_05min
    - hs_29to33_15min
    - hs_29to33_30min
2. heatshock with sorbitol
    - 29C(1M_sorbitol)~33C(1M_sorbitol)_05min
    - 29C(1M_sorbitol)~33C(1M_sorbitol)_15min
    - 29C(1M_sorbitol)~33C(1M_sorbitol)_30min
3. heatshock from 25°C to 37°C
    - hs_05min_hs-1
    - hs_10min_hs-1
    - hs_15min_hs-1
    - hs_30min_hs-1
    - hs_40min_hs-1
    - hs_60min_hs-1
    - hs_80min_hs-1
4. hypothermia from 37°C to 25°C
    - hs_37to25_15min
    - hs_37to25_30min
    - hs_37to25_45min
    - hs_37to25_60min
    - hs_37to25_90min
5. sorbitol treatment
    - 1M_sorbitol_05min
    - 1M_sorbitol_15min
    - 1M_sorbitol_30min
    - 1M_sorbitol_45min
    - 1M_sorbitol_60min
    - 1M_sorbitol_90min
    - 1M_sorbitol_120min
6. nitrogen depletion
    - Nitrogen_Depletion_30min
    - Nitrogen_Depletion_1h
    - Nitrogen_Depletion_2h
    - Nitrogen_Depletion_4h
    - Nitrogen_Depletion_8h
    - Nitrogen_Depletion_12h
    - Nitrogen_Depletion_1d
7. diamide treatment

- 1.5mM_diamide_5min
- 1.5mM_diamide_10min
- 1.5mM_diamide_20min
- 1.5mM_diamide_30min
- 1.5mM_diamide_40min
- 1.5mM_diamide_50min
- 1.5mM_diamide_60min

8. dithiothrietol exposure
    - 1.5mM_diamide_5min
    - 1.5mM_diamide_10min
    - 1.5mM_diamide_20min
    - 1.5mM_diamide_30min
    - 1.5mM_diamide_40min
    - 1.5mM_diamide_50min
    - 1.5mM_diamide_60min
9. menadione exposure
    - 1mM_Menadione_10min_redo
    - 1mM_Menadione_20min_redo
    - 1mM_Menadione_30min_redo
    - 1mM_Menadione_40min_redo
    - 1mM_Menadione_50min_redo
10. hydrogen peroxide treatment
    - 0.32mM_H2O2_10min_redo
    - 0.32mM_H2O2_20min_redo
    - 0.32mM_H2O2_30min_redo
    - 0.32mM_H2O2_40min_rescan
    - 0.32mM_H2O2_50min_redo
    - 0.32mM_H2O2_60min_redo

Bibliography

1. Alberts, B., Molecular biology of the cell. 1989: Second edition. New York : Garland Pub., [1989] ©1989.
2. Hentze, M.W., et al., A brave new world of RNA-binding proteins. Nature Reviews Molecular Cell Biology, 2018. 19(5): p. 327-341.
3. Kwasnieski, J.C., et al., Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proceedings of the National Academy of Sciences, 2012. 109(47): p. 19498.
4. Moore, M.J., From Birth to Death: The Complex Lives of Eukaryotic mRNAs. Science, 2005. 309(5740): p. 1514.
5. Mayr, C., Regulation by 3'-Untranslated Regions. Annual Review of Genetics, 2017. 51(1): p. 171-194.
6. Mignone, F., et al., Untranslated regions of mRNAs. Genome biology, 2002. 3(3): p. REVIEWS0004-REVIEWS0004.
7. Szostak, E. and F. Gebauer, Translational control by 3'-UTR-binding proteins. Briefings in Functional Genomics, 2012. 12(1): p. 58-65.
8. Wilkie, G.S., K.S. Dickson, and N.K. Gray, Regulation of mRNA translation by 5'- and 3'-UTR-binding factors. Trends in Biochemical Sciences, 2003. 28(4): p. 182-188.
9. Gerber, A.P., D. Herschlag, and P.O. Brown, Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. PLoS Biol, 2004. 2(3): p. E79.
10. Ulbricht, R.J. and W.M. Olivas, Puf1p acts in combination with other yeast Puf proteins to control mRNA stability. RNA (New York, N.Y.), 2008. 14(2): p. 246-262.
11. Rajagopal, N., et al., High-throughput mapping of regulatory DNA. Nature Biotechnology, 2016. 34(2): p. 167-174.
12. Shlyueva, D., G. Stampfel, and A. Stark, Transcriptional enhancers: from properties to genome-wide predictions. Nature Reviews Genetics, 2014. 15(4): p. 272-286.
13. Suryamohan, K. and M.S. Halfon, Identifying transcriptional cis-regulatory modules in animal genomes. Wiley Interdiscip Rev Dev Biol, 2015. 4(2): p. 59-84.
14. Vijayabaskar, M.S., et al., Identification of gene specific cis-regulatory elements during differentiation of mouse embryonic stem cells: An integrative approach using high-throughput datasets. PLoS Comput Biol, 2019. 15(11): p. e1007337.
15. Wissink, E.M., E.A. Fogarty, and A. Grimson, High-throughput discovery of post-transcriptional cis-regulatory elements. BMC Genomics, 2016. 17(1): p. 177.
16. Shen, Y., et al., A map of the cis-regulatory sequences in the mouse genome. Nature, 2012. 488(7409): p. 116-120.
17. Muller, F., P. Blader, and U. Strahle, Search for enhancers: teleost models in comparative genomic and transgenic analysis of cis regulatory elements. Bioessays, 2002. 24(6): p. 564-72.
18. Chen, W.J. and T. Zhu, Networks of transcription factors with roles in environmental stress response. Trends in Plant Science, 2004. 9(12): p. 591-596.
19. Castello, A., M.W. Hentze, and T. Preiss, Metabolic Enzymes Enjoying New Partnerships as RNA-Binding Proteins. Trends in Endocrinology & Metabolism, 2015. 26(12): p. 746-757.

20. Yuan, M. and Y. Lin, Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006. 68(1): p. 49-67.
21. Gasch, A.P., et al., Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell, 2000. 11(12): p. 4241-57.
22. Shalgi, R., et al., A catalog of stability-associated sequence elements in 3' UTRs of yeast mRNAs. Genome biology, 2005. 6(10): p. R86-R86.
23. Hogan, D.J., et al., Diverse RNA-Binding Proteins Interact with Functionally Related Sets of RNAs, Suggesting an Extensive Regulatory System. PLOS Biology, 2008. 6(10): p. e255.
24. Cheng, J., et al., Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast. RNA (New York, N.Y.), 2017. 23(11): p. 1648-1659.
25. Troyanskaya, O., et al., Missing value estimation methods for DNA microarrays. Bioinformatics, 2001. 17(6): p. 520-525.
26. Friedman J, Hastie T, Tibshirani R (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, **33**(1), 1–22. http://www.jstatsoft.org/v33/i01/.
27. Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.
28. Abhishek Jain, 2019 (Undergraduate Honours Project, unpublished)
29. Wikipedia contributors. Messenger RNA. *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 15 Apr. 2020. Web. 16 Apr. 2020. https://en.wikipedia.org/wiki/Messenger_RNA
30. Yevgeniy, G. Introduction to DNA Microarrays. *bitesizebio*, 15 Apr. 2020. Web. 16 Apr. 2020. https://bitesizebio.com/7206/introduction-to-dna-microarrays/
31. Google Classroom. Overview of transcription. Khan Academy, 15 Apr. 2020. Web. 16 Apr. 2020. https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/overview-of-transcription
32. Daniel P. Riordan, Daniel Herschlag, Patrick O. Brown, Identification of RNA recognition elements in the *Saccharomyces cerevisiae* transcriptome, *Nucleic Acids Research*, Volume 39, Issue 4, 1 March 2011, Pages 1501–1509, https://doi.org/10.1093/nar/gkq920
33. Wang, Ming et al. The PUF Protein Family: Overview on PUF RNA Targets, Biological Functions, and Post Transcriptional Regulation. *International journal of molecular sciences* vol. 19,2 410. 30 Jan. 2018, doi:10.3390/ijms19020410