Bachelor thesis
Computer Science

Radboud University

Inferring gene regulatory relationships from gene expression data

Author:
T.A. (Tom) van Bussel
s4221435

First supervisor/assessor:
Dr. ir. Tom Claassen
tomc@cs.ru.nl

Second assessor:
Dr. Tjeerd Dijkstra
t.dijkstra@science.ru.nl

April 13, 2016

Abstract

In order to understand how genes affect each others expression, we want to infer regulatory relationships between genes and use these genes to build gene regulatory networks. Several algorithms exist for inferring regulatory relationships between genes. One of the state of the art algorithms is Trigger, but Trigger seems to produce unsatisfactorily high probability estimates in some cases. In this thesis we analyze several issues of Trigger which are related to the estimation of the local false discovery rate, which Trigger uses to calculate its probability estimates. We show that even though Trigger is able to identify regulation relationships, the issues lead to an overestimation of the probabilities for regulation relationships.

We will also demonstrate that a new approach using Bayes factors for correlation matrices can be applied to this problem and does not suffer from these issues. We apply the new algorithm, which we call BFCM, to an experiment in yeast in order to show that it is able to produce rich and biologically coherent information about the underlying gene regulatory relationships. The new algorithm produces more conservative probability estimates than Trigger, and is able to identify new regulation relationships.

Contents

Chapter 1

Introduction

Large scale gene expression and genotype data has become abundant since the rise of microarrays [6, 30, 10]. Microarrays have allowed us to measure genetic variation, and RNA and protein expression levels for thousands of genes on hundreds of individuals [31, 19, 20]. This made quantative trait locus (QTL) mapping possible, which is a method that identifies genetic regions that are linked to a phenotype of interested, such as the expression level of gene. The QTL mapping does not give full information about the interaction between genetic locations and gene expression levels as genes influence each other through regulation. In order to fully understand how genes affect each others expression levels, the "wiring diagram" is of great interest, which describes how genes regulate each other and how they interact.

Inferring gene regulatory networks is of great interest, as it can help with studying complex diseases such as cancer. An example of such research, is the research by Emmert-Streib et al. [15] in which they infer a gene regulatory network of breast cancer, which they use to identify genes that play a role in breast cancer. Figure 1.2 shows the gene regulatory network which they found.

Recently a large number of algorithms have been published that try to infer these regulation relationships between genes [3, 17, 18, 8, 21, 22]. In thesis we will look at one of these algorithms, Trigger [8]. Trigger constructs the underlying regulation network by estimating the probabilities of one gene regulating another. Trigger has several issues however; it is known to overestimate some of these probabilities [7] and it is unable to detect regulation relationship which also include hidden variables which affect both of the genes [8].

In this thesis we analyze the issues of Trigger that cause probability estimates that seem to be too high and we determine the causes of these errors.

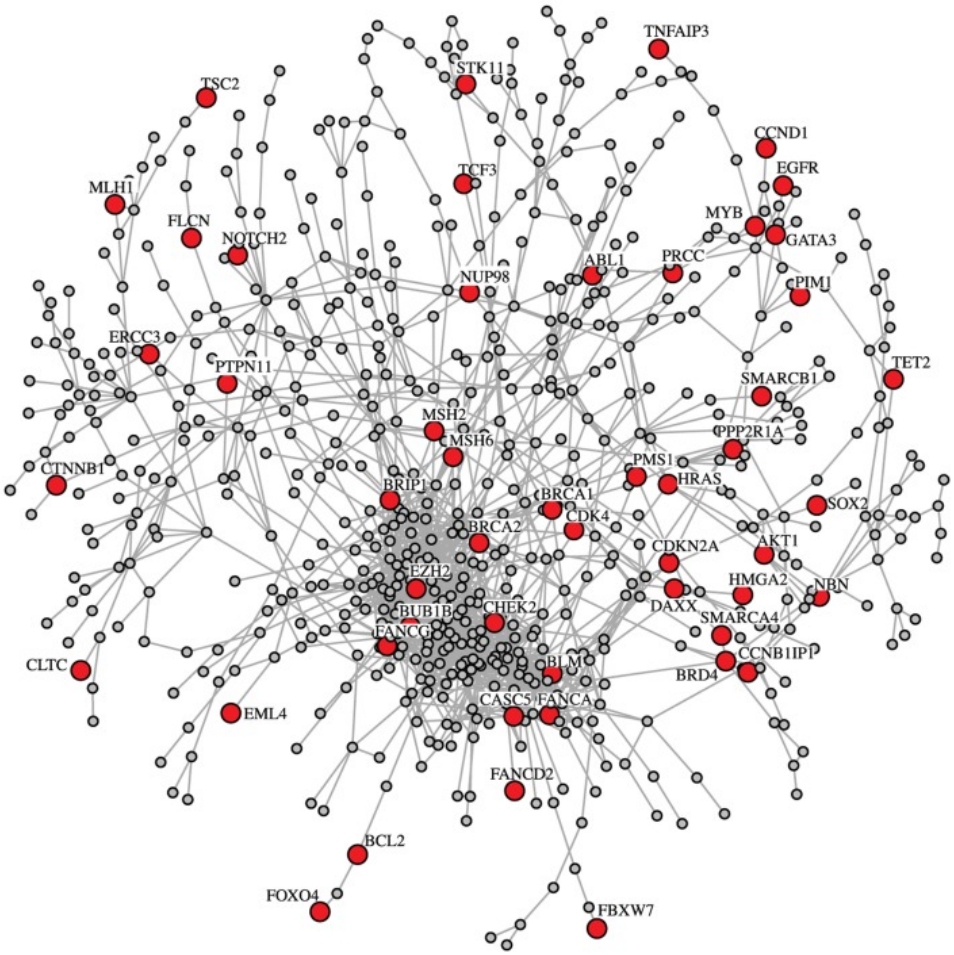

Figure 1.1: Gene regulatory network of breast cancer. Taken from Emmert-Streib et al. [15].

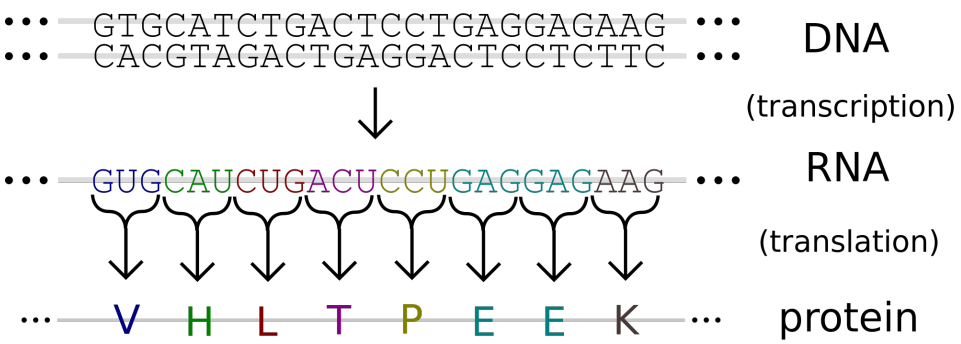

Figure 1.2: Gene expression. Taken from Wikipedia.

We will also try to find solutions for these problems. Finally we introduce a new algorithm for the inferring of gene regulation relationships, which we call BFCM and uses correlation matrices and Bayes factors [16]. We will show that this algorithm is able to produce rich and biologically coherent information by applying it on a gene expression dataset for yeast [6, 5] and we will compares these results to the results of Trigger..

The remaining part of this chapter provides a short overview of the biology behind gene expression and putative regulation relationships between gene and provides an overview of the remaining chapters in this thesis.

1.1 Genomics

Cells store their hereditary information in the form of DNA. In order to carry its information-bearing function DNA must express its information. During a process called gene expression the information stored on DNA is used the guide the creation of other molecules in the cell. Gene expression consists of two main steps: first the gene is transcribed to a molecule called messenger RNA. This messenger RNA is then transported outside the cell's nucleus and is translated into long sequences of 20 different animo acids called proteins by a complex molecule machine called a ribosome.

Proteins can influence how cells function, and proteins can also influence the process of gene expression. Through regulation proteins can influence both the transcription of DNA and the translation of messenger RNA. Not only can a protein influence the expression of the gene that it was copied from, but it can also regulate the gene expression of different genes. This process is called gene regulation and allows a cell to express proteins when needed, which improves the versatility and adaptability of an organism.

A Quantitative Trait Locus (QTL) is a section of DNA, which is called the locus, that highly correlates with the variation of a quantitative trait. Especially of interest are Expression QTLs (eQTLs) which are genomic loci which contribute to a variation of expression levels of messenger RNAs and proteins. We make a distinction between two types of linkage between genes and QTL. The first type is called cis-linkage in which eQTLs are mapped to a location on the DNA which is close to the gene of origin. The second type of linkage is trans-linkage in which the expression levels of proteins and mRNA are linked to loci that are far way from their gene of origin.

We refer the reader to "Molecular Biology of the Cell" by Alberts et al. [1] for more information about gene expression.

In order to analyze regulation between genes data about the loci and expression levels of genes have to collected. Two inbred lines, individuals who are completely homozygous, having identical alleles of each gene on both homologous chromosomes, by repeated sibling mating, are crossed to create new individuals with randomized DNA. The simplest cross is a backcross, in which two inbred lines are crossed to obtain the first filial generation. As this generation receives a copy of every chromosome of each of the two parental strains, the first filial generation is heterozygous, thus having different alleles of some genes on both homologous chromosomes. These individuals are then crossed with one of the two parental strains, which results in a new generation with a chromosome that is a mosaic of the chromosome of the two parental strains.

A more complicated approach uses recombinant inbred lines. This approach again crosses the two parental strains to create a first filial generation, but instead of creating a backcross with one of the two parental strains, an individual from the first filial generation is crossed with another individual from that generation two create N pairs of individuals in the second filial generation. Finally these pairs are crossed in recombinant inbred lines until we have N individuals that are completely homozygous at every locus.

Once we have N segregants we can collect the necessary data for analysis. The individuals are genotyped to find their genetic structure. Microarrays are used to identify genetic markers on the genome and to measure the alleles on each of the N individuals. The allele is used to identify from which parental strain each piece of DNA comes from. Microarrays are also used to measure the expression levels of each gene by measuring the activity of proteins and messenger RNA corresponding to those genes. These two measurements give us the allele at each genetic marker for each individual and the expression level of each gene for each individual.

1.2 Thesis organization

The following chapter provides a description of Trigger [8] and we show how the local false discovery rate is used to calculate probability estimates for gene regulation relationships. The chapter also provides a short analysis of the yeast dataset using Trigger. The third chapter analyzes two issues that cause Trigger to produce estimates that are higher than expected.

The fourth chapter introduces a new algorithm for inferring gene regulatory relationships, called BFCM. The algorithm applies theory based on Bayes factors of correlation matrices to produce a probability estimate for gene regulation relationships. The remainder of the chapter analyzes the yeast dataset using BFCM and compares the results with this results of the analysis using Trigger.

And finally the final chapter will provide a conclusion and will discuss some of the future work we plan to do.

Chapter 2

Trigger Algorithm

Trigger (Transcriptional Regulation Inference from Genetics of Gene ExpRession) [8] is an algorithm that can be used to reconstruct transcriptional regulatory networks and to identify putative regulators of genes.

2.1 Description of the algorithm

Trigger estimates the probability P_{ij} that the transcription of gene i is a regulator of the transcription of any other gene j, after which the user has to decide for which probabilities the regulation relationship is significant. Trigger calculates conservative estimates of these probabilities which are denoted by $\hat{P}_{ij}$. Using these estimated probabilities Trigger constructs a directed graph which is used to represent the regulatory network in which nodes represent genes and directed edges represent regulation relations between genes. If the estimated probability, $\hat{P}_{ij}$, is higher than an user set threshold λ, a directed edge between genes i and j is added in the graph.

In order to calculate these probabilities we first have to define what it mathematically means that gene i regulates gene j. The regulation of gene i by gene j corresponds to the causal model $L \rightarrow T_i \rightarrow T_j$. Here L is used to denote the locus, T_i is used to denote the transcription level of gene i and T_j is used to denote the transcription level of gene j. We are interested whether or not L causes T_i and T_i causes T_j. By the causation $T_i \rightarrow T_j$ we mean that a causal manipulation of T_i will change the distribution of T_j, but an ideal manipulation of T_j will not cause a change in the distribution of T_i.

The locus L is used to determine causation. The locus L is properly randomized, as its randomization takes place before the expression levels of the T_i's are measured. Thus the association of L and the expression level of a T_i implies a causation of T_i by L.

In order to estimate the probability P_{ij} Trigger estimates the probability $\Pr(L \to T_i \to T_j)$, the probability that the causal model $L \to T_i \to T_j$ explains the gene expression data. As this model is quite complicated, it is difficult to directly estimate this probability. The Causal Equivalence Theorem is used to split the model in three simpler models. The theorem states that the causal relationship $L \to T_i \to T_j$ exists and there are no hidden variables causal for both T_i and T_j if and only if the conditions $L \to T_i$, $L \to T_j$ and $L \perp T_j | T_i$ hold. The conditions $L \to T_i$ and $L \to T_j$ ensure that T_i and T_j are both randomized by L and the condition $L \perp T_j | T_i$ ensure that the causal effect from L on T_j is fully explained by T_i.

By splitting the causal model in three simpler conditions we can also split the probability $\Pr(L \to T_i \to T_j)$ in three probabilities that are much easier to estimate:

$$\begin{aligned}
&\Pr(L \to T_i \to T_j) \\
&= \Pr(L \to T_i \text{ and } L \to T_j \text{ and } L \perp T_j | T_i) \\
&= \Pr(L \to T_i) \Pr(L \to T_j | L \to T_i) \Pr(L \perp T_j | T_i | L \to T_i, L \to T_j)
\end{aligned}$$

As a consequence of the usage of the Causal Equivalence Theorem Trigger is unable to detect regulation relationships in which there is confounding caused by hidden variables. Trigger detects the cases when there are no confounding hidden variables, and is only able to calculate correct probabilities in these cases. Chen et al. claim that this causes Trigger to produce conservative estimates of P_{ij}, as $\Pr(T_i \to T_j) \geq \Pr(L \to T_i \to T_j) \geq \Pr(L \to T_i \to T_j$ and there exists no hidden variable H such that $H \to T_i$ and $H \to T_j)$, the probability of regulation relationships with confounding hidden variables are underestimated, but this does not necessarily imply that the probability estimates are conservative, as the estimate itself can still overestimate P_{ij}.

Algorithm 1 provides an algorithmic description of Trigger. Trigger first finds the marker with strongest local linkage to T_i for each gene g and then estimates the probabilities for each of the three models. The next section explains how each of the probabilities are estimated.

2.2 Estimation of probabilities

The estimation of the three probabilities is split in five steps. First the expression data, $t_{i1}, t_{i2}, \ldots, t_{in}$, for each gene is transformed to follow a standard normal distribution using the following formula:

$$t^*_{ik} = \Phi^{-1}\left(\frac{\text{rank}(t_{ik})}{n+1}\right) \qquad k = 1, 2, \ldots, n$$

Algorithm 1 Algorithmic description of Trigger

```
p ← LOC_LINK_P_VALUES
for all genes g do
    loc_markers ← GET_LOCAL_MARKERS(marker.pos, exp.pos, g)
    l[g] ← MIN(p[g, local_markers])
    loc_prob[g] ← CALC_LOC_PROB(l, p, g)
end for
for all genes g1 do
    for all genes g2 do
        sec_prob[g1, g2] ← CALC_SEC_PROB(l, g1, g2)
        ind_prob[g1, g2] ← CALC_IND_PROB(l, g1, g2)
    end for
end for
return loc_prob × sec_prob × ind_prob
```

The next three steps involve the generation of the observed and null statistics, the latter are generated using permutation testing. First we have to generate the statistics for the primary linkage, $L \to T_i$. The model

$$t_{ik} = \alpha_i + \beta_i \ell_k + \epsilon_{ik}$$

can be used to test if T_i is linked to L_i. Under the null hypothesis of no linkage $\beta_i = 0$ and under the alternative hypothesis of linkage $\beta_i \neq 0$. Using a permutation test the null statistics X_i^{0b} are generated by replacing t_{ik} with $t_{i,r(j)}$, where r is a random permutation of $1, 2, \ldots, n$.

The next step is the calculation of the test statistics for the secondary linkage, $L \to T_j$. Again we model the relationship between T_j and L as $t_{jk} = \alpha_j + \beta_j \ell_k + \epsilon_{jk}$, but now we need to consider that there is a linkage between T_i and L as the test is conditioned on $L \to T_i$. As (T_i, T_j) jointly follow a bivariate normal distribution, the two variables have the following distribution when conditioned on L:

$$\begin{pmatrix} t_{ik}|\ell_k \\ t_{jk}|\ell_k \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha_i + \beta_i \ell_k \\ \alpha_j + \beta_j \ell_k \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \sigma_{ij} \\ \sigma_{ij} & \sigma_j^2 \end{pmatrix} \right)$$

Again, under the null hypothesis of no linkage $\beta_j = 0$ and under the alternative hypothesis of linkage $\beta_j \neq 0$. Just like when we tested for primary linkage, we have to compute an observed likelihood ratio statistic Y_{ij} and a number of permuted statistics Y_{ij}^{0b}. These null statistics are computed by permuting the expression data of T_j. Note that in the previous step we could have also permuted the genotype markers ℓ_k, but in this step this is no longer as it would also remove the primary linkage which we conditioned on.

In the fourth step the observed and null statistics for the conditional independence between L and T_j given T_i are calculated. Contrary to the previous two tests we will now have to test for independence instead of testing for dependence. As this is proves to be much more difficult we will instead test for dependence. A test of $T_j|T_i$ and L being independent is equivalent to a test of $(T_j - \rho_{ij}T_i)|T_i$ and L being independent, where ρ_{ij} is the correlation between T_i and T_j. Under the null hypothesis the distribution of $(t_{jk} - \rho_{ij}t_{ik})|t_{ik}, \ell_k$ will not depend on ℓ_k and thus it will be distributed by a single normal distribution (with mean zero and a variance of $1 - \rho_{ij}^2$). Under the alternative hypothesis the distribution will depend on the allele and thus $(t_{jk} - \rho_{ij}t_{ik})|t_{ik}, \ell_k$ will be distributed by a mixture of normal distributions with unspecified allele-specific means and variances. In order to calculate the null statistics the expression data is permuted for both gene i and gene j with seperate permutations.

In the final step the calculated statistics are used to compute empirical Bayesian estimates for the probabilities. The local false discovery rate[12, 13, 14], which is the posterior probability that null hypothesis is true given a test statistic X. The local false discovery rate is defined as

$$\text{fdr}(X) = \frac{\pi_0 f_0(X)}{\pi_0 f_0(X) + (1 - \pi_0) f_1(X)},$$

with $f(X) = \pi_0 f_0(X) + (1 - \pi_0) f_1(X)$ the mixture density of the statistics X, where f_0 is the null density, the density function of the statistics for which the null hypothesis is true, and f_1 is the alternative density, and π_0 the proportion of true null hypotheses. When using p values the local false discovery rate can be simplified as $\text{fdr}(p) = \frac{\pi_0}{f(p)}$ with the simplified mixture density $f(p) = \pi_0 + (1 - \pi_0) f_1(p)$, as independent p values that come from the null hypothesis have a uniform distribution between 0 and 1.

The local false discovery rate is used to provide a probability estimate for each of the three models. In order to estimate the probability of the primary linkage, $L \to T_i$, we can simply estimated the local true discovery rate, which is one minus the local false discovery rate, over all i.

$$\Pr(L \to T_i | X_i) = 1 - \text{fdr}(X_i)$$

We can use the same strategy for each fixed i to estimate the probability of secondary linkage:

$$\Pr(L \to T_j | L \to T_i, Y_{ij}) = 1 - \text{fdr}(Y_{ij})$$

The same strategy cannot be applied to estimate the probability of conditional independence for a twofold of reasons: we now want to calculate the

probability that the null hypothesis is true and we need to condition on primary and secondary linkage. In order to properly condition on primary and secondary linkage we select the $(1 - \pi_0^{iY})$ most significant proportion of T_j for secondary linkage from the previous step and only their corresponding Z_{ij} are used for the estimation of the local false discovery rate:

$$\Pr(L \perp T_j | T_i | L \to T_i, L \to T_j, Z_{ij}) = \text{fdr}(Z_{ij})$$

Finally the three probabilities are multiplied to form the estimate of P_{ij}:

$$\begin{aligned}&\Pr(L \to T_i \to T_j) = \\ &\quad \Pr(L \to T_i) \times \Pr(L \to T_j | L \to T_i) \times \Pr(L \perp T_j | T_i | L \to T_i, L \to T_j)\end{aligned}$$

2.3 Implementation details

In this section we will discuss a number of issues that arise during the implementation of Trigger.

In order to detect causation Trigger not only looks at T_i and T_j, the transcripts for genes i and j, but also at a locus L. This requires us to consider each triplet (L, T_i, T_j) instead of each pair (T_i, T_j). In the case of the yeast dataset this means that we now have to consider 125,323,635,360 triplets instead of just 38,632,440 pairs. In order to improve the computational efficiency, Trigger only considers one locus, L_i, for each gene i, which is the locus with strongest cis-linkage to transcript i.

The second issue that arrises is that we need to select a method for the estimation of the local false discovery rate. A large number of these methods involve the estimation of the densities of the observed p values under the null and alternative hypotheses, the considered methods differ in the way that they estimate these densities. As p values have an uniform distribution under the null hypothesis, we only have estimate the density of the p values under the alternative hypothesis. Chen et al. [9] use kernel density estimation for all three probabilities, but our implementation we will only kernel density estimation for primary and secondary linkage and we use a Beta-Uniform Model (BUM) for the conditional independence. In the next section we will argue that kernel density estimation is unsuited when trying to determine conditional independence between L and T_j given T_i.

It is possible to approximate the distribution of the null statistics using analytic distributions. As Trigger uses likelihood ratio tests we can approximate the distribution of two times the log of the likelihood ratio statistic with a Chi-squared distribution. As this is only an approximation this will reduce the accuracy of the estimated probabilities, but as we are no longer

required to compute the permuted statistics the algorithm becomes significantly more efficient. We've opted to use an approximation for the primary and secondary linkage, but to use empirical p values for the conditional independence.

2.4 Results

We used Trigger to analyze an experiment of yeast [6, 5].

In this experiment two strains of genes were crossed to produce 112 independent recombinant segregants. The cross involved the two parental strains BY4716, which is isogenic to the laboratory strain S288C, and the wild isolate RM11-1a, which was acquired from a Califnornia vineyard. The expression levels of 6216 were measured and genotypes were measured on 3244 markers, which covered 99% of the genome. The genetic map of the marker locations is shown in figure 2.1.

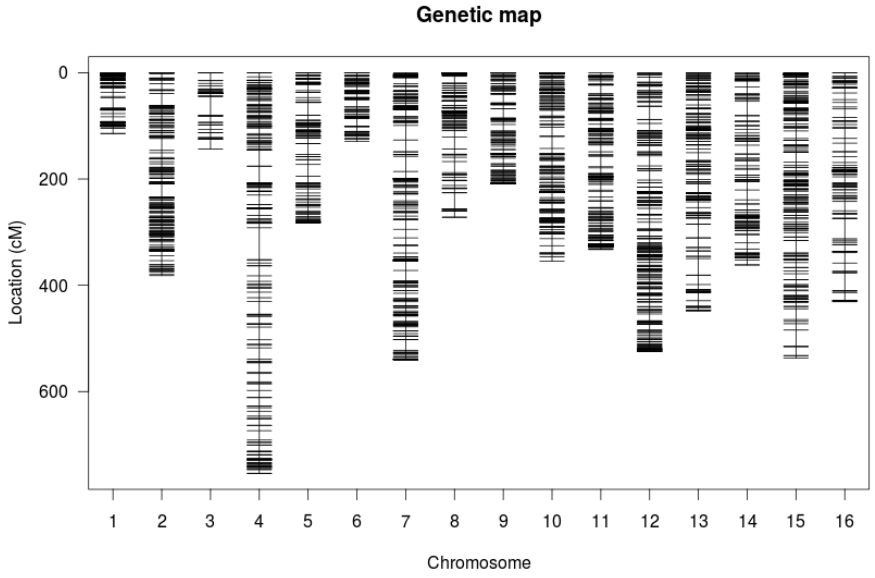

Figure 2.1: Genetic map of the 3244 marker locations in the yeast dataset

We examined two genes in detail, both of which are suspected regulators. NAM9 is a suspected regulator [8] which is located on chromosome 14 and is a component of mitochondrion [27]. NAM9 also is a structural constituent of ribosome, it is involved in mitochondrial translation and the mitochondrial ribosomal small subunit [29].

ILV6 is another suspected regulator [8] which is located on chromosome 3 and is a regulatory subunit of acetolactate synthase, which catalyzes the first step of branched-chain amino acid biosynthesis, enhances the activity of the Ilv2p catalytic subunit and localizes to mitochondria [11, 24].

We used Trigger to analyze linkages between gene transcripts and loci. In order to find locally linked genes we performed likelihood ratio tests for all markers within a 50 kb window of the gene. We limited the locus to be in a 50 kilobase region of the transcript T_i in order to increase statistical and computational efficiency, this region was large enough such that most genes were linked to a locus in their region. Figure 2.2 shows the locations of the markers and genes on the genome that are linked. We used a p value cut-off corresponding to a FDR of 5%, using the methodology described in [32]. The figure shows large amount of cis-linkage, which is indicated by the diagonal line. Vertical lines in the figure indicate linkage hotspots, both chromosome 3 (the chromosome of ILV6) and chromosome 14 (where NAM9 is located) seem to be linkage hotspots.

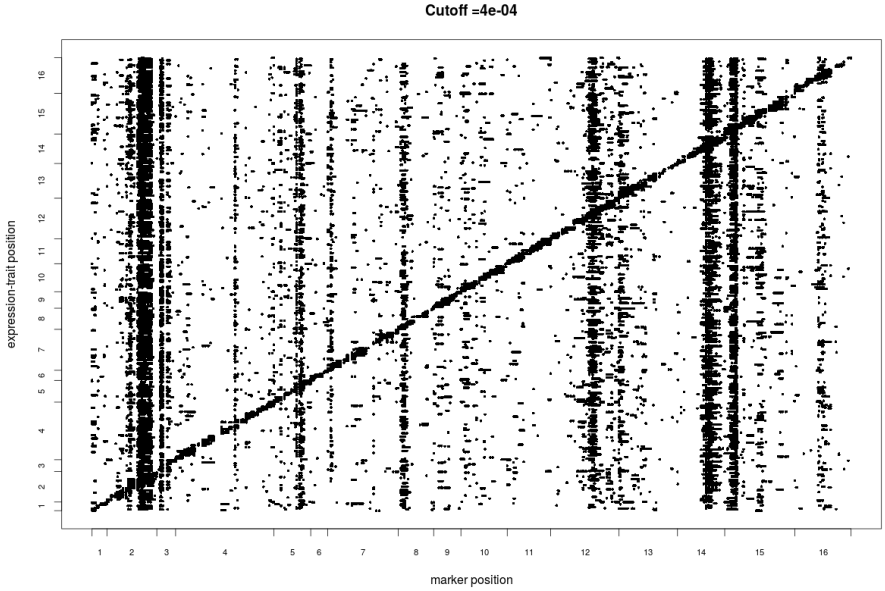

Figure 2.2: Genome-wide eQTL and gene expression linkage map.

We also used Trigger to construct a gene regulatory network for the experiment of yeast. We compared our implementation with the results reported by Chen et al. Figure 2.1 shows the results for genes significantly regulated by NAM9, showing large differences in the reported probabilities. Figure 2.2 shows the results for genes significantly regulated by ILV6,

Gene	Rank (Chen et al.)	Prob (Chen et al.)	Prob (Ours)
MDM35	1	0.973	0.828
CBP6	2	0.969	0.827
QRI5	3	0.960	0.805
RSM18	4	0.959	0.816
RSM7	5	0.954	0.818
MRPL11	6	0.925	0.778
MRPL25	7	0.888	0.750
DLD2	8	0.872	0.759
YPR126C	9	0.861	0.730
MSS116	10	0.849	0.752

Table 2.1: Genes regulated by NAM9 with estimated regulation as estimated by Chen et al. and by our implementation.

showing more similar results. The differences in probabilities indicate that estimations by Trigger are unstable.

Gene	Rank (Chen et al.)	Prob (Chen et al.)	Prob (Ours)
TRP4	1	0.999	1.000
ARG2	2	0.988	0.991
YPL264C	3	0.977	0.981
GGC1	4	0.951	1.000
LYS4	5	0.948	0.980
NPR1	6	0.947	0.954
ASN1	7	0.938	0.969
CCP1	8	0.937	0.956
YKR015C	9	0.928	0.938
CPA2	10	0.928	0.988

Table 2.2: Genes regulated by ILV6 with estimated regulation as estimated by Chen et al. and by our implementation.

Chapter 3

Analysis of Trigger

This chapter provides an analysis of two issues in Trigger which lead to the overestimation of the regulation probabilities. The previous chapter showed that the probability estimate for a regulation relationship was split in three parts, in the case of local linkage and secondary linkage the probability is defined as one minus the local false discovery rate and in the case of the conditional independence it is defined as simply the local false discovery rate. The next sections show that this problematic in the case of conditional independence, due to two issues that can cause the local false discovery rate to overestimate. The next section explains how Trigger estimates the local false discovery rate, and the remaining sections discuss its issues.

3.1 Estimation of the local false discovery rate

Trigger estimates the probabilities using the local false discovery for p values, which for p values is defined as

$$\text{fdr}(p) = \frac{\pi_0}{f(p)},$$

where π_0 is the proportion of true null hypotheses, and f is the mixture density, which is defined as

$$f(p) = \pi_0 + (1 - \pi_0) f_1(p),$$

where f_1 is the density of the p values for which the alternative hypothesis is true. The proportion of true null hypotheses π_0 and the mixture density f have to be estimated in order to estimate the local false discovery rate.

In order to estimate the local false discovery rate we first have to estimate the proportion of p values that come from the null hypothesis, π_0. This is difficult without knowing the distribution of alternative p values, but we can use the fact that null p values are uniformly distributed between zero

and one, as the p value is the probability that a more extreme statistic is found under the null hypothesis, to find a reasonable estimate. Storey and Tibshirani [32] introduced the following estimation method, which is used by Trigger [9]:

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, \ldots, m\}}{m(1-\lambda)}$$

which involves a tuning parameter λ. The method assumes that p values larger than λ the null hypothesis is true and that they are thus uniformly distributed. This means that the height of the region $p > \lambda$ can be viewed as a conservative estimate of π_0, in the sense that it does not underestimate the proportion of true null hypotheses.

The tuning parameter has to be set such that the region $p > \lambda$ is mostly flat, which would indicate that most of the p values in that region are uniformly distributed. In order to do this automatically, Storey and Tibshirani estimate π_0 for a range of value of λ and then fit a natural cubic spline with three degrees of freedom through the estimated values to approximate π_0 as a function of λ. Finally this cubic spline is estimated in $p = 1$ and the resulting value is used as the estimate for π_0. Figure 3.1 shows a visual description of this method.

The other component that needs to be estimated is the mixture density f, which is defined as

$$f(p) = \pi_0 + (1 - \pi_0) f_1(p)$$

and is the mixture of null density (which is uniform in the case of p-values) and the alternative density function, f_1, with mixing weight π_0. Trigger estimates the mixture density as a whole, instead of plugging in the estimate for π_0 and estimating the alternative density f_1. Trigger uses Kernel Density Estimation [28, 25] and in order to deal with the bounded support of p-values it transforms [33] the p-values using the quantile function of a standard normal distribution. Figure 3.2 shows how the mixture density function and the local false discovery are estimated after transformation.

3.2 Issue 1: the estimate of π_0 is an upper bound

The first issue occurs during the estimation of the proportion of true null hypotheses, π_0. The estimation method for π_0 assume that the mixture density of the p values, f, is decreasing and that $\pi_0 = f(1)$. However these assumptions are not always correct. As $\pi_0 \leq f(p) = \pi_0 + (1 - \pi_0) f(p)$ for all p, this will lead to an overestimation of the true π_0, which will lead to an overestimation of the local false discovery rate which is equal to $\pi_0 / f(p)$,

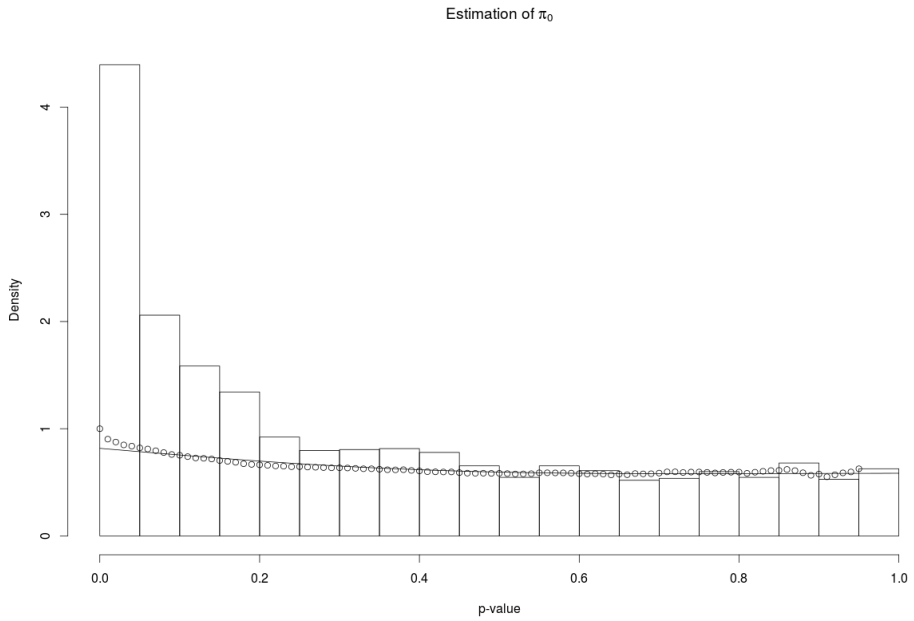

Figure 3.1: Estimation of π_0. The histogram gives a rough estimate of density of the p values, the dots represent the estimates $\hat{\pi}_0(\lambda)$, and the line is the cubic spline used to estimate π_0.

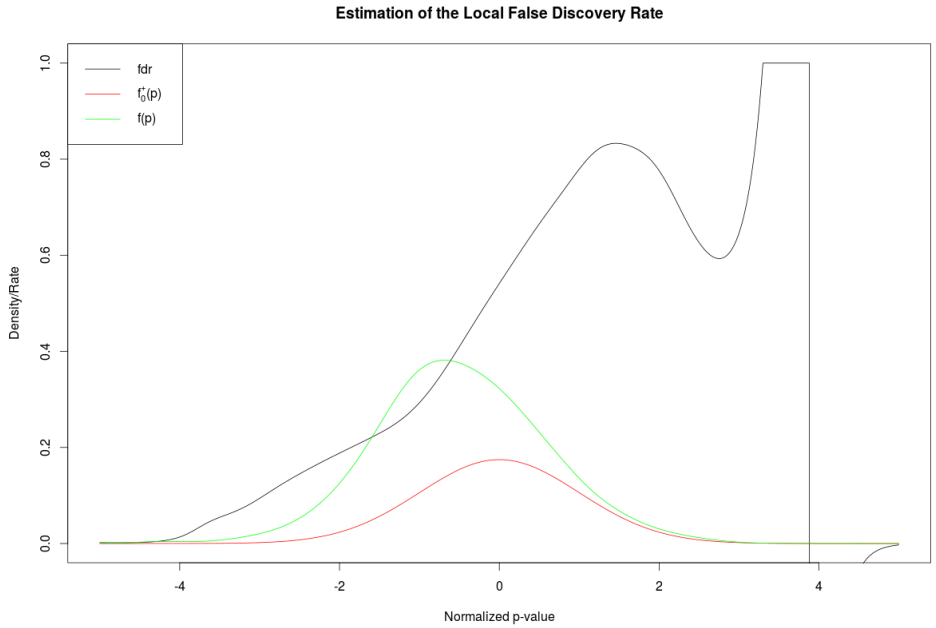

Figure 3.2: Estimation of the fdr for conditional independence for regulator NAM9.

and thus it will lead to an overestimation of the probability for conditional independence between L_i and T_j given T_i.

Nguyen and Matais [23] show that π_0 and f_1 are identifiable on a set $(0,1) \times F$ if and only if for all $g_1 \in \mathcal{F}$ and for all $c \in (0,1)$ we have that $c+(1-c)g_1 \notin \mathcal{F}$. Here $\mathcal{F}$ is the set of possible alternative density functions. This is not necessarily the case when using a likelihood ratio test, so π_0 is not guaranteed to be identifiable.

In the case of Trigger, which uses two-sided likelihood ratio tests, the distributions of the alternative p values are unknown, thus π_0 is not always identifiable. When π_0 is unidentifiable $\hat{\pi}_0$ serves as an estimated upper bound for π_0. This causes no issues in the case of local linkage and secondary linkage as $\Pr(L \to T_i|p) = 1-\text{fdr}(p)$. In this case we find an estimated upper bound of the fdr and thus the estimated probability is conservative. In the case of conditional independence however the probability is overestimated, as in this case the probability is equal to the fdr, which is overestimated.

During the estimation of π_0 it is assumed that the mixed density function of the p values is decreasing and that thus $f(1) \leq f(p)$ for all $p \in [0,1]$. As the mixture density function is defined as $f(p) = \pi_0 + (1-\pi_0)f_1(p)$, π_0 is smaller or equal than $f(p)$ for all $p \in [0,1]$, and thus $\pi_0 \leq \min_{p\in[0,1]} f(p)$. If we incorrectly assume that f is always the lowest in $p = 1$, a better upper bound might be available. This means that Trigger possibly overestimate the local false discovery and thus the probability of conditional independence even more than necessary.

3.3 Issue 2: estimation of π_0 and f are decoupled

Trigger decouples the estimation of the proportion of true null hypotheses and the mixture density. Careful estimation of the mixture density is required to ensure that $\hat{f}(p) \geq \hat{\pi}_0$ for all p, otherwise the estimate of the local false discovery rate $\hat{\pi}_0/\hat{f}(p)$ will be larger than 1, which causes the estimated probability to be above 1 in the case of conditional independence (in the case of local and secondary linkage this causes no issues as we are not interested in low probabilities). We have observed this behaviour during the analysis of the yeast dataset using Trigger. Figure 3.3 shows the estimated local false discovery rate for the conditional indepence with ILV6 and shows that for high p values the local false discovery rate, an thus the probability, could be estimated as exactly 1.

Figure 3.3 shows that for p values larger than 0.7 the local false discovery rate and thus the estimated probabilities of conditional indepence are

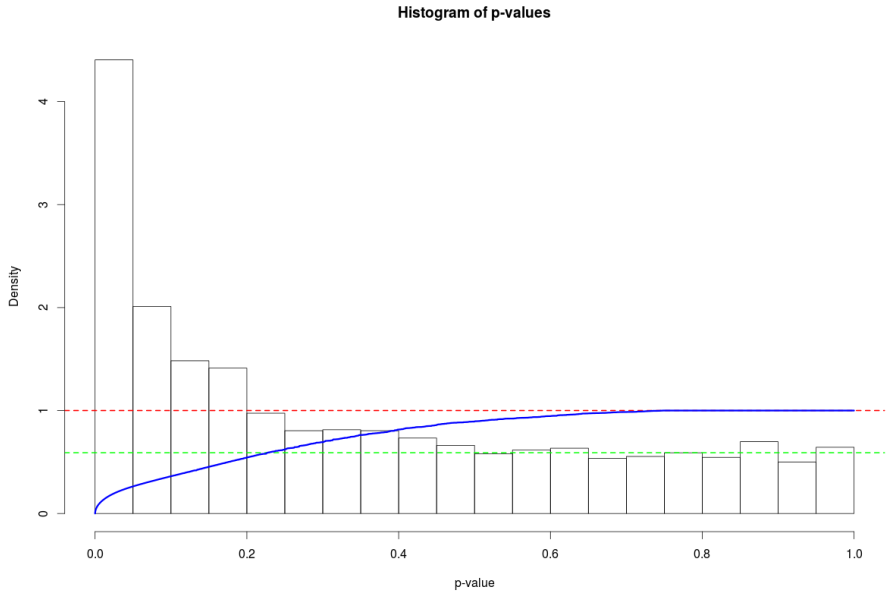

Figure 3.3: Plot of the fdr for conditional independence for the regulator ILV6. Note that the local false discovery is estimated as equal to 1 for p values above 0.8

close or equal to one. The histogram of the p values, which provides a simple approximation of the density function of the p values, suggests that the empirical density between 0.65 and 0.75 is below $\hat{\pi}_0$, which would cause a local false discovery rate above one and hence after cut-off of the local false discovery rate an estimated probability of conditional independence that is larger than one.

Figure 3.4 shows a plot of the estimated false discovery rates, but this time we did not enforce the monotonicity of local false discovery rate for p values and we did not cut-off values above one. It shows that between p values 0.6 and 0.85 the estimated local false discovery rate is larger than one. This clearly indicates that the estimate of the local false discovery rate is wrong, as the local false discovery rate is the probability that the null hypothesis is true for a given p value and thus cannot be higher than one.

Figure 3.5 shows that similar errors occur during the estimation of the local false discovery rate for secondary linkage. In this case however these kinds of errors are less troublesome as these errors occur for high p values and produce a high fdr and in the case of secondary linkage we are interested in a low fdr for low p values.

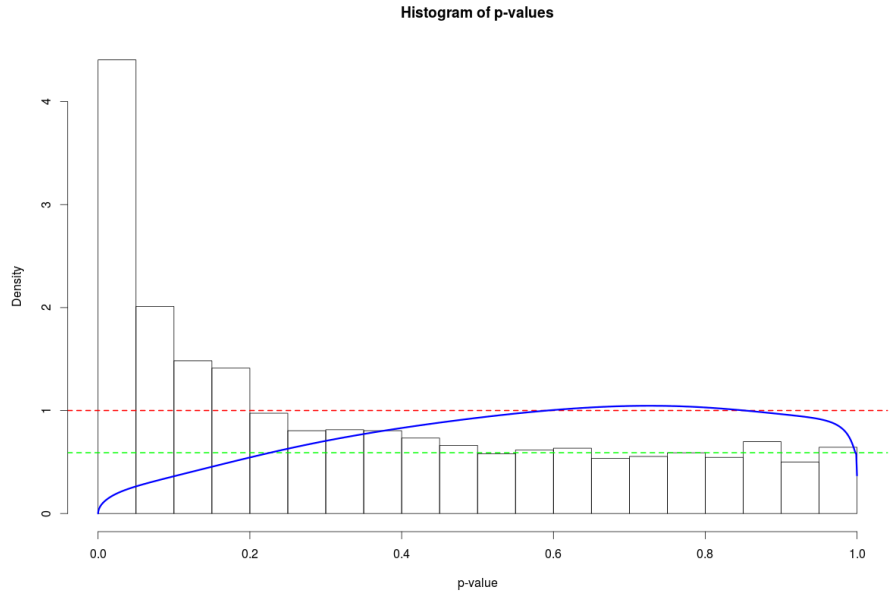

Figure 3.4: The local false disocery rate estimate for conditional independence for the regulator ILV6. The fdr is no longer corrected for monotonicity and fdr values above one are no longer cut off. It is now visible that the fdr incorrectly gets estimated for p values between 0.6 and 0.8.

These errors occur because the estimation of π_0 and f are decoupled. And thus it is never enforced that $\pi_0 \leq f(p)$ for all p-values p and that thus the estimate $\hat{\pi}_0$ should be smaller than the estimate $\hat{f}(p)$ for all p-values p.

3.3.1 Solutions

In order to solve this issue the estimation of π_0 and f should be coupled: either f should be estimated while ensuring that $\hat{f}(p) \geq \hat{\pi}_0$ for all p values p, or π_0 should be estimated such that $\hat{\pi}_0 \leq \min_{p\in[0,1]} f(p)$.

Efron et al. [14] noted that this behavior should not happen and used this to property to find the following upper bound for π_0 for p values:

$$\pi_0 \leq \min_{p\in[0,1]} f(p)$$

Using this property they found the following estimator for π_0:

$$\hat{\pi}_0 = \min_{p\in[0,1]} \hat{f}(p).$$

Note that the minimum of the ratio of the estimated densities no longer forms an upper bound for $\hat{\pi}_0$, but that when $\hat{\pi}_0 \geq \min_{p\in[0,1]} \hat{f}(p)$ errors will

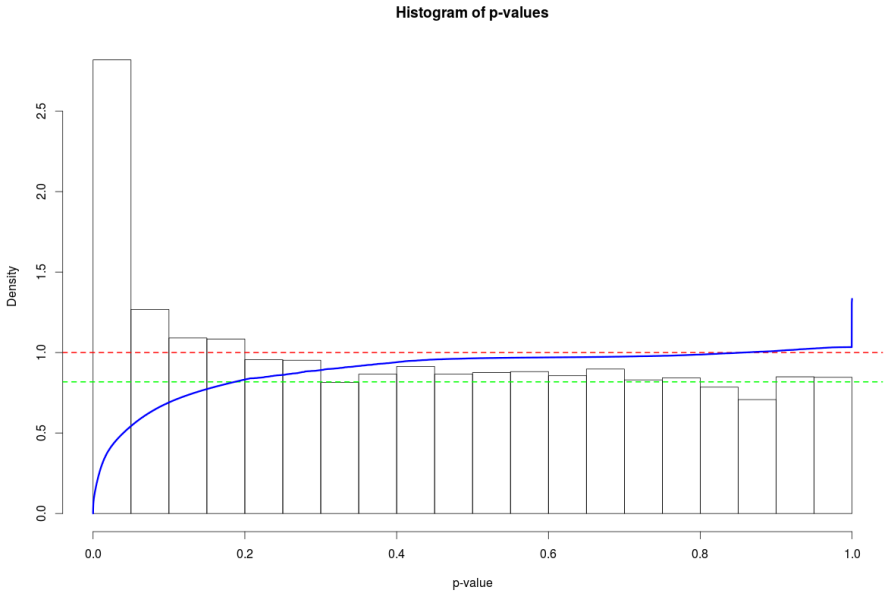

Figure 3.5: The fdr for secondary linkage for regulator NAM9.

occur in the estimation of the local false discovery rate. Efron et al. note that much better estimates for π_0 are available.

In order to improve the estimation of the mixture density without decreasing the quality of the estimation of the proportion of true null hypotheses, a number of assumptions are made [23]

(a) f is monotonically decreasing.

(b) $\lim_{p \to 1} f(p) = \pi_0$.

(c) $f \geq \pi_0$ for all p values.

As shown in the previous section the second assumption is incorrect, but as this assumption was already implicitly made by Trigger we believe that this will not worsen the estimation of the local false discovery rate.

These assumptions allow us to estimate the mixture density using a Beta-Uniform Mixture model [26]. This method fits a number of Beta distributions to observed p-values.

$$f(x) = \lambda + (1 - \lambda) a x^{a-1}$$

The parameters λ and a are estimated using Maximum Likelihood estimation. Pounds and Morris use this model to estimate π_0 using $\hat{\pi}_0 =$

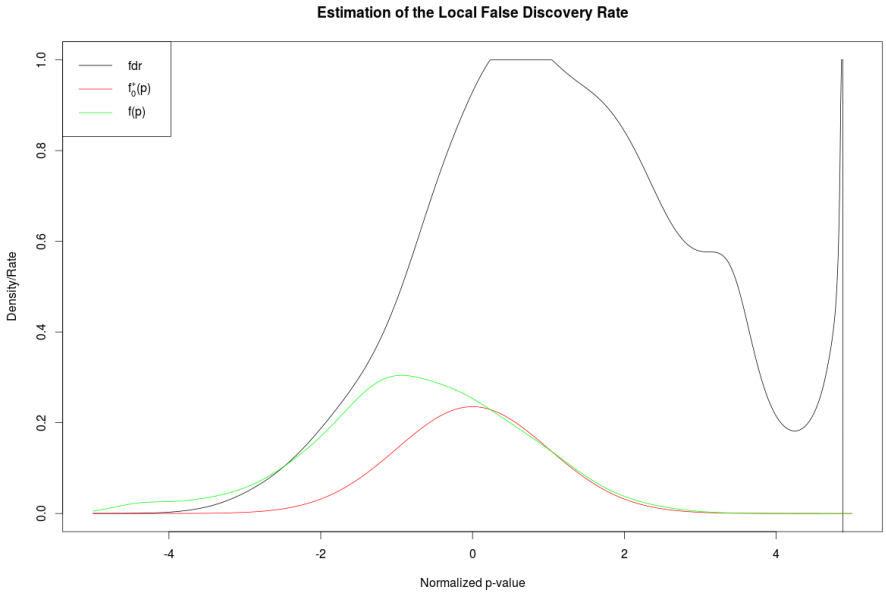

Figure 3.6: Estimation of the fdr for conditional independence for regulator ILV6. Note that the mixture density f (green line) is lower than $\pi_0 \cdot f_0$ (red line), leading to a local false discovery estimate of one (black line).

$\hat{\lambda} + (1 - \hat{\lambda})\hat{a}$, but instead we will use our estimate π_0 to estimate the mixture density. We fit the following more general distribution to the observed p-values:

$$f(p) = \pi_0 + (1 - \pi_0)\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}$$

We plug in our estimate of π_0 and we use maximum likelihood estimation to estimate α and β, we restrict α to values in $(0, 1)$ and β to values in $(1, \infty)$ to ensure that the mixture density function is non-increasing.

Figure 3.7 shows the results when using the BUM model for the estimation of the mixture density and the local false discovery rate. Compared 3.3 which uses a kernel density estimation which is decoupled from the estimation of π_0 there no longer is a region of p values for which the local false discovery rate is equal to one. As the BUM method explicitly assumes that f is non-increasing there no longer is a need to adjust the local false discovery rate for monotonicity.

3.4 Summary of the issues

In this chapter we analyzed the seemingly high probability estimates that Trigger calculates. We found two issues in Trigger that could explain these

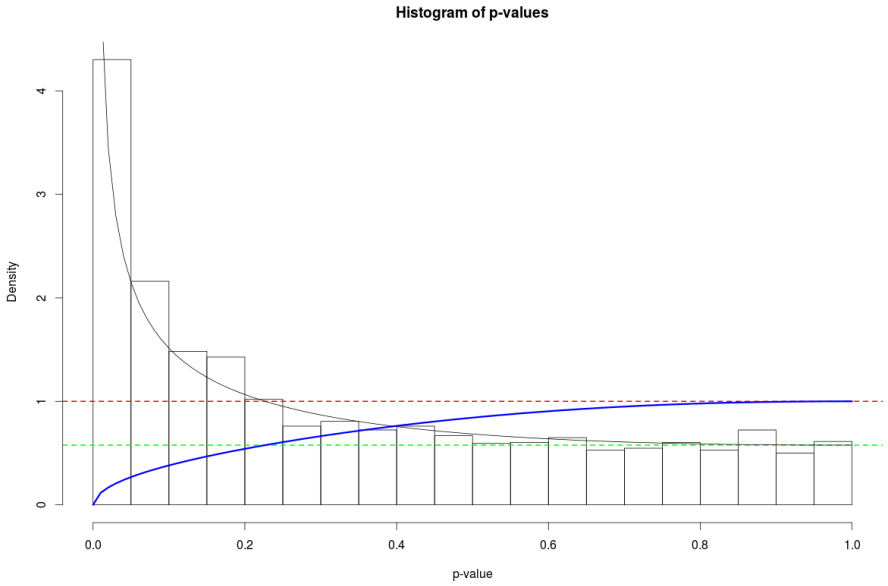

Figure 3.7: Estimation of the fdr for conditional independence for regulator ILV6 using a BUM model.

high estimates. Both of these issues occur during the estimation of the local false discovery rate, which Trigger uses as the probability for conditional independence between L_i and T_j given T_i. The first issue is caused by the incorrect assumption that for high p values the null hypotheses is always true. This causes an overestimation of the proportion of true null hypotheses, which in turn causes an overestimation of the local false discovery rate. The second issue is caused by the decoupled estimation of f and π_0. This allows situations to occur in which $\hat{\pi}_0 > \hat{f}(p)$ which causes the estimate of the local false discovery rate, $\hat{\pi}_0/\hat{f}(p)$ to be larger than one.

Chapter 4

Bayes Factors of Correlation Matrices

In the previous section we have shown two issues that cause Trigger to overestimate the probabilities for regulatory relationships between genes, due to problems in the estimation of the local false discovery rate. In this section we will show how a method based on a Bayesian factors using correlation matrices [16] can be applied infer regulatory relationships between gene transcripts. This approach completely avoids the usage of local false discovery rates by scoring a larger number of models and thus does not suffer from the same issues as Trigger.

4.1 Description of the algorithm

The goal of BFCM is to infer the structure between three stochastic variables, X_1, X_2 and X_3. Five types of structures are considered by BFCM: full independence ($X_1 \perp X_2 \perp X_3$), acausal single independence ($X_1 \perp X_2$), causal conditional independence ($X_1 \perp X_2 | X_3$), one independent variable ($X_1 \perp (X_2, X_3)$) and full dependence ($X_1 \not\perp X_2 \not\perp X_3$). Algorithm 2 gives a short overview of the algorithm.

Similarly to Trigger first selects the marker with the highest local linkage for each gene transcript. Instead of calculating a test statistic BFCM measures the absolute value of the correlation between markers and expression levels and selects the locus in a region near the gene with the highest absolute correlation to the expression levels. As the genotype markers are randomized before the expression levels a measured, a correlation between the genotype markers and the expression levels of the gene transcripts implies a causation of the expression levels by the genotype markers.

Once the number of triples (L, T_i, T_j) has been reduced, we calculate the

Algorithm 2 Algorithmic description of BFCM

```
corr_1 ← CORRELATION(marker, exp)
for all genes g do
    loc_markers ← GET_LOCAL_MARKERS(marker.pos, exp.pos, g)
    l[g] ← MAX(corr[g, local_markers])
end for
corr_2 ← CORRELATION(exp, exp)
for all genes g_1 do
    for all genes g_2 do
        Pr[g_1, g_2] ← SCORE_CORR_PATTERNS(corr_1, corr_2, l, g_1, g_2)
    end for
end for
return Pr
```

probability of a structure explaining the data:

$$\Pr(S|D) = \frac{P(D|S)\Pr(S)}{P(D)}$$

. By conditioning on S we can rewrite $P(D)$ in terms of $P(D|S)$ and $P(S)$:

$$P(D) = \sum_S P(D|S)\Pr(S)$$

These probabilities are estimated using Bayes factors of correlation matrices [16]. The correlation matrices are transformed to covariance matrices and a Bayes factor comparing a structure S_1 with another structure S_2:

$$K = \frac{\Pr(D|S_1)}{\Pr(D|S_2)} = \frac{\int \Pr(\Sigma|S_1)\Pr(D|\Sigma, S_1)d\Sigma}{\int \Pr(\Sigma|S_2)\Pr(D|\Sigma, S_2)d\Sigma}.$$

The algorithm constructs a correlation matrix for the (T_i, T_j) pair and then for each triplet it constructs a 3×3 correlation matrix, ρ, which is computed by extracting the correlation coefficients from the two larger correlation matrix. For each (L, T_i, T_j) triple we calculate a Bayes factor for each structure, in which compare the structure with the structure $X_1 \not\perp X_2 \not\perp X_3$:

$$K(X_1 \perp X_2 \perp X_3) = c_1(n, v)c_{\frac{1}{2}}(n, v)\|\rho\|^{\frac{n+v}{2}}$$

$$K(X_1 \perp (X_2, X_3)) = c_1(n, v)\left(\frac{\|\rho\|}{1-\rho_{23}^2}\right)^{\frac{n+v}{2}}$$

$$K(X_1 \perp X_2 | X_3) = c_{\frac{1}{2}}(n, v)\left(\frac{\|\rho\|(1-\rho_{12}^2)}{(1-\rho_{12}^2)(1-\rho_{13}^2)(1-\rho_{23}^2)}\right)^{\frac{n+v}{2}}$$

$$K(X_1 \perp X_2) = \frac{c_1(n, v)}{c_{\frac{1}{2}}(n, v)}(1-\rho_{12}^2)^{\frac{n+v-1}{2}}$$

with:

$$c_1(n,v) = \frac{n+v-2}{v-2}$$
$$c_{\frac{1}{2}}(n,v) = \frac{\Gamma(\frac{n+v}{2})\Gamma(\frac{v-1}{2})}{\Gamma(\frac{n+v-1}{2})\Gamma(\frac{v}{2})}$$

We also need to choose a prior on the eleven structures. One option is to represent the causal relationships using directed acyclic graphs (DAGs). We then count the number of DAGs for each possible structure and compare that to the total number of DAGs. Figure 4.1 shows the structures which are considered by BFCM in order to infer the regulatory relationships between genes. Some of the possible DAGs are missing from this list of structures as these are biologically impossible, such as $T_i \rightarrow L_i$.

Using the Bayes factors we can then estimate the probability of a structure:

$$\Pr(S|D) = \frac{K(S)\Pr(S)}{\sum_T K(T)\Pr(T)} \tag{4.1}$$

in which the common term $1/\Pr(D|X_1 \not\perp X_2 \not\perp X_3)$ cancels out. The estimated probability $\Pr(L_i \perp T_j | T_i | D)$ is then used as probability estimate for the regulation relationship $L_i \rightarrow T_i \rightarrow T_j$.

4.2 Results

We analyzed the yeast dataset using BFCM, in order to analyze the algorithm. Table 4.1 shows the number number of putative regulators found and the number of edges found for different probability thresholds. For instance for a threshold of 0.75 we found 5042 significant regulatory relationships among 2580 genes of which 365 were regulators, these edges have a false discovery rate of 24.1%. Figure 4.2 shows the distribution of the probabilities found by BFCM. No probabilities above 80% were found, even though this can be unsatisfactory, it also means that the estimated probabilities are more conservative than those found by Trigger which means that the FDR won't be underestimated as much.

These results show that BFCM estimates are much more conservative than the estimates found by Trigger. The probability estimates found by BFCM also have a slower fall-off than Trigger and they highest probability is at roughly 80%. This means that it could be harder to select a suitable probability cut-off using the False Discovery Rate.

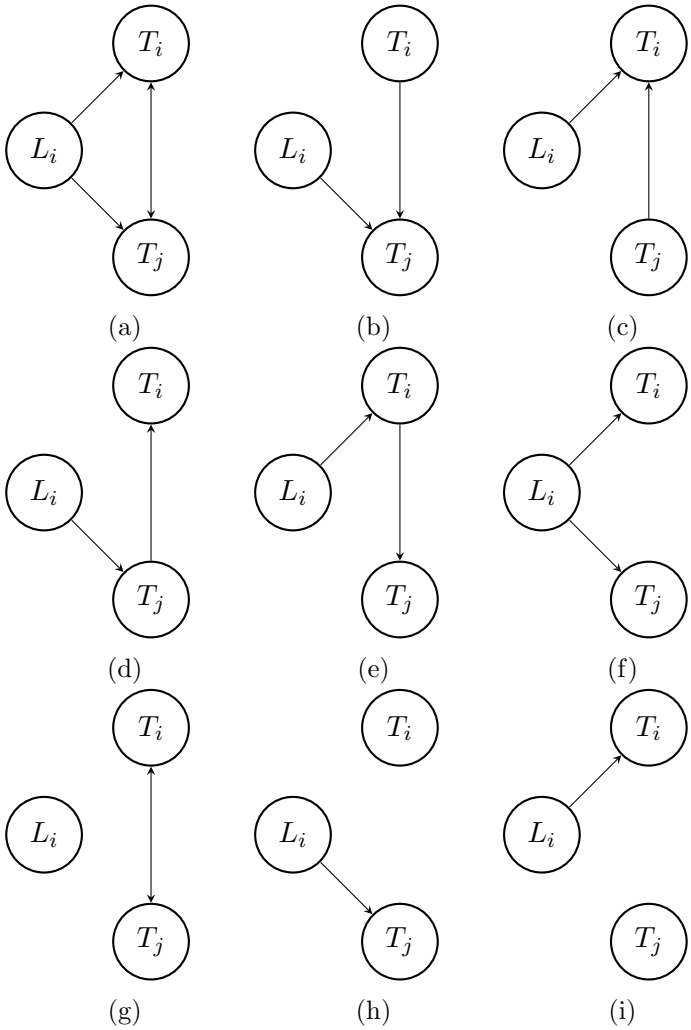

Figure 4.1: Models considered by BFCM. (a) Full dependence. (b) Single independence ($L_i \perp T_i$). (c) Single independence ($L_i \perp T_j$). (d) Conditional Independence ($L_i \perp T_i | T_j$). (e) Conditional independence ($L_i \perp T_j | T_i$). (f) Conditional independence ($T_i \perp T_j | L_i$). (g) One independent variable ($L_i \perp (T_i, T_j)$). (h) One independent variable ($T_i \perp (L_i, T_j)$). (i) One independent variable ($T_j \perp (L_i, T_i)$). (j) Full independence (not pictured). Arrows with two directions indicate that there two possible structures for the model.

Probability	Number of putative regulators	Total number of genes	Number of edges	FDR (%)
0.77	12	39	27	22.9
0.76	209	1488	2002	23.5
0.75	365	2580	5042	24.1
0.7	851	5266	31531	27.2

Table 4.1: Number of putatitive regulators and regulation relationships found by BFCM at different probability cut-offs.

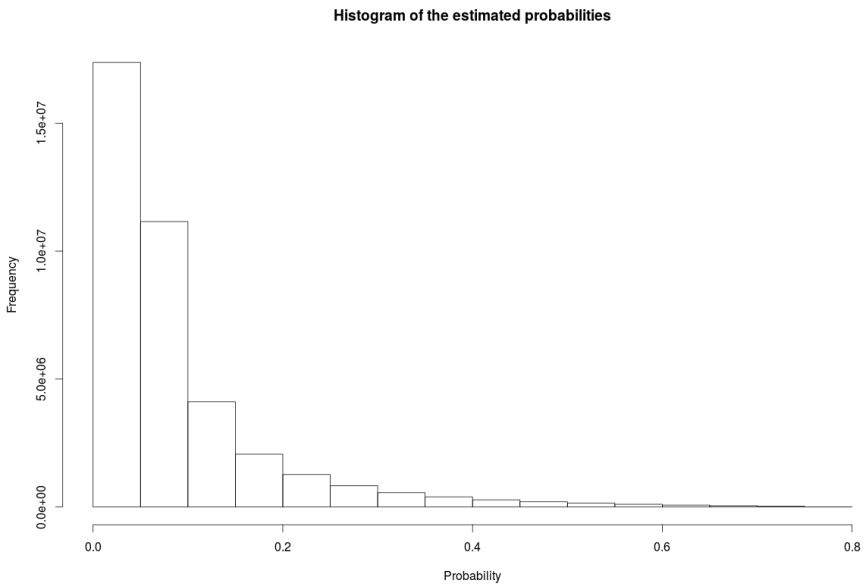

Figure 4.2: Histogram of probability estimates by BFCM. The histogram shows that most of the gene pairs get assigned a low probability estimate.

Gene	Rank (BFCM)	Prob (BFCM)	Rank (Trigger)	Prob (Trigger)
MRPS8	1	0.764	87	0.473
UBP16	2	0.763	257	0.318
MTO1	3	0.763	510	0.209
MRP1	4	0.763	706	0.154
IFM1	5	0.763	738	0.148
COX10	6	0.763	631	0.173
MRPS5	7	0.762	58	0.528
DIA4	8	0.762	134	0.402
MDM35	9	0.761	2	0.766
MNP1	10	0.761	146	0.391

Table 4.2: Genes regulated by NAM9 with highest probabilities for BFCM and corresponding rank and probabilities for Trigger

Gene	Rank (Trigger)	Prob (Trigger)	Rank (BFCM)	Prob (BFCM)
CBP6	1	0.768	13	0.759
MDM35	2	0.766	9	0.761
RSM7	3	0.756	14	0.758
RSM18	4	0.747	18	0.754
QRI5	5	0.742	39	0.743
MRPL11	6	0.720	20	0.754
MSS116	7	0.696	27	0.750
AFG3	8	0.694	28	0.749
DLD2	9	0.693	23	0.752
MRPL25	10	0.688	62	0.729

Table 4.3: Genes regulated by NAM9 with highest probabilities for Trigger and corresponding rank and probabilities for BFCM

Gene	Rank (BFCM)	Prob (BFCM)	Rank (Trigger)	Prob (Trigger)
QDR3	1	0.769	1	1.0
TRP4	2	0.769	3	1.0
ARG1	3	0.768	1134	0.386
YDR476C	4	0.768	97	0.890
ARG5	5	0.767	553	0.589
LAP3	6	0.766	1649	0.216
HIS1	7	0.766	92	0.894
BSC5	8	0.765	235	0.782
ARG2	9	0.765	15	0.989
YIL056W	10	0.763	48	0.938

Table 4.4: Genes regulated by ILV6 with highest probabilities for BFCM and corresponding rank and probabilities for Trigger

Gene	Rank (Trigger)	Prob (Trigger)	Rank (BFCM)	Prob (BFCM)
QDR3	1	1.0	1	0.769
GGC1	2	1.0	13	0.762
TRP4	3	1.0	2	0.769
CPA2	4	1.0	24	0.756
FRE6	5	1.0	35	0.751
YPR059C	6	1.0	29	0.753
ADE3	7	0.998	37	0.750
DMA2	8	0.998	65	0.735
UGA3	9	0.997	56	0.742
HIS5	10	0.996	41	0.749

Table 4.5: Genes regulated by ILV6 with highest probabilities for Trigger and corresponding rank and probabilities for BFCM

GO term	P value	Cluster frequency	Background frequency	FDR (%)	Genes
mitochondrial translation	9.325e-22	18 out of 26	170 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MNP1, MRPL3, MRP49, RSM7, MRPL27, MRPL11
mitochondrion organization	9.836e-22	22 out of 26	429 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MTO1, MDM35, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, MRPL11
single-organism biosynthetic process	1.946e-11	20 out of 26	934 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MNP1, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
translation	1.538e-10	18 out of 26	753 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MNP1, MRPL3, MRP49, RSM7, MRPL27, MRPL11
peptide biosynthetic process	1.606e-10	18 out of 26	757 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MNP1, MRPL3, MRP49, RSM7, MRPL27, MRPL11
peptide metabolic process	3.375e-10	18 out of 26	798 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MNP1, MRPL3, MRP49, RSM7, MRPL27, MRPL11
amide biosynthetic process	4.983e-10	18 out of 26	806 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MNP1, MRPL3, MRP49, RSM7, MRPL27, MRPL11
cellular amide metabolic process	1.551e-9	18 out of 26	861 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MNP1, MRPL3, MRP49, RSM7, MRPL27, MRPL11
organelle organization	3.584e-9	22 out of 26	1622 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MTO1, MDM35, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, MRPL11
single-organism organelle organization	1.138e-8	19 out of 26	1127 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MTO1, MNP1, MRPL3, MRP49, MRPL27, RSM7, MRPL11
organonitrogen compound biosynthetic process	1.465e-8	19 out of 26	1143 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MNP1, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
organonitrogen compound metabolic process	3.601e-7	19 out of 26	1369 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MNP1, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
single-organism metabolic process	0.000001067	22 out of 26	2133 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MTO1, MNP1, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
cellular protein metabolic process	1.306e-6	20 out of 26	1680 out of 7166	0.00%	CBP6, UBP16, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, YNR020C, MNP1, MRPL3, MRP49, MRPL27, RSM7, MRPL11
cellular component organization	2.048e-6	22 out of 26	2202 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MTO1, MDM35, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, MRPL11
protein metabolic process	4.346e-6	20 out of 26	1794 out of 7166	0.00%	CBP6, UBP16, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, YNR020C, MNP1, MRPL3, MRP49, MRPL27, RSM7, MRPL11
cellular component organization or biogenesis	4.530e-5	22 out of 26	2566 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MTO1, MDM35, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, MRPL11
cellular nitrogen compound biosynthetic process	4.854e-5	19 out of 26	1819 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MNP1, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
gene expression	6.640e-5	20 out of 26	2087 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MTO1, YNR020C, MNP1, MRPL3, MRP49, MRPL27, RSM7, MRPL11
cellular macromolecule biosynthetic process	0.0004719	18 out of 26	1854 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MNP1, MRPL3, MRP49, RSM7, MRPL27, MRPL11
macromolecule biosynthetic process	0.0005172	18 out of 26	1865 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MNP1, MRPL3, MRP49, RSM7, MRPL27, MRPL11
organic substance biosynthetic process	0.0006894	20 out of 26	2386 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MNP1, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
biosynthetic process	0.0008792	20 out of 26	2418 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MNP1, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
cellular biosynthetic process	0.003231	19 out of 26	2348 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MNP1, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
cellular metabolic process	0.004386	24 out of 26	3934 out of 7166	0.00%	CBP6, UBP16, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MTO1, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
organic substance metabolic process	0.006802	24 out of 26	4014 out of 7166	0.00%	CBP6, UBP16, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MTO1, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
cellular macromolecule metabolic process	0.009326	21 out of 26	3066 out of 7166	0.00%	CBP6, UBP16, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MTO1, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, MRPL11
cellular nitrogen compound metabolic process	0.009512	20 out of 26	2783 out of 7166	0.00%	CBP6, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, MRPS17, MRP1, DIA4, NAM9, MRPS8, MTO1, MNP1, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11

Table 4.6: GO Terms for biological processes for NAM9 and top 25 regulated genes found by BFCM

Just as in previous sections we looked at two genes in detail: NAM9 on chromosome 14 and ILV6 on chromosome 3. In section 3.1 we saw that both genes had high cis-linkage and each locus that they were locally linked to showed large amounts of trans-linkage to other genes. At a 75% probability cut-off BFCM found 27 significantly regulated genes by NAM9 and 38 significantly regulated genes by ILV6. Tables 4.2, 4.3, 4.4 and 4.5 show the genes with the 10 highest probability estimates for BFCM and Trigger, for NAM9 and ILV6. These results show that many of genes that get a high probability estimate from Trigger also receive a high probability estimate from BFCM, but many of genes that received a high probability estimate from BFCM did not receive a high probability estimate from Trigger.

In order to determine whether significantly regulated genes were related to their regulator we used the Gene Ontology (GO) database [2]. We employed the tool GO Term Finder [4] to find the significant terms among regulators and regulated genes. This approach infers information from separately and independently performed experiments and allowed us to test specifically whether common processes, functions, and components are present among each set of genes. We found that all regulated genes were significantly related to their regulators.

Tables 4.6, A.1 and A.2 show the significant GO terms for NAM9 and its significantly regulated genes. All genes that were found to be significantly regulated by NAM9 using BFCM share significant GO terms with NAM9. Furthermore all but two of the terms found annotate NAM9, which further suggests that NAM9 is a putative regulator for these genes. NAM9 isn't the only gene that appears in all but two of the terms though, other such genes include MRPS8 and RSM18.

Tables A.3 and A.4 show the significant GO terms for ILV6 and its significantly regulated genes. All genes that were found to be significantly

regulated by ILV6 using BFCM share significant GO terms with ILV6. Furthermore all but three of the terms found annotate ILV6, which further suggests that ILV6 is a putative regulator for these genes. ILV6 isn't the only gene that appears in all but two of the terms though, other such genes include HOM3 and ARG2.

Chapter 5

Conclusions

In this thesis, we investigated algorithms which can be used to infer regulatory networks between genes. We investigated Trigger, which seemed to produce probability estimates that are too high, and we developed a new approach, which we call BFCM, to infer causal regulatory relationships among genes, which is based on Bayes factors of correlation matrices.

We have shown that Trigger overestimates some of the probabilities for regulation relationships among genes, and that these estimates are unstable. After thresholding the probabilities, these overestimated probabilities can be incorrectly identified as regulation relationships in the gene regulatory network. These issues are caused by an overestimation of the local false discovery rate which is used as the probability estimate when testing for conditional independence between L_i and T_j given T_i in the model $L_i \to T_i \to T_j$, and thus leads to an overstimation of the probabilities.

The first issue occurs during the estimation of the proportion of true null hypotheses, π_0, and its identifiability. We have shown that currently available estimates for this proportion are unsuited as they give an upper bound for π_0, which results in an upper bound for the local false discovery rate.

The second issue is caused by the decoupled estimation of the proportion of true null hypotheses and the mixture density of the p values. We have shown that in the case the value of the mixture density f is close to $\hat{\pi}_0$, the estimated upper bound of π_0, kernel density estimation can provide an estimate of $f(p)$ which causes an overestimation of the local false discovery rate.

We believe that Trigger is suitable for the discovery of putative regulation relationships, even though Trigger overesimates the probabilities, as gene regulation relationships that get assigned a high probability estimate

by Trigger also get a high probability estimate from BFCM, which avoids issues in the estimation of the local false discovery rate by using Bayes factors instead.

We introduced a new algorithm for inferring gene regulatory network called BFCM, which does not suffer from these issues as it uses Bayes factors instead of local false discovery rates in its estimates. We have demonstrated how BFCM can be applied to inferring causal regulatory relationships between genes. We applied BFCM to an experiment in yeast in which 112 recombinant lines were monitored for genome-wide expression. Using analysis of the Gene Ontology database of two suspected putative regulators we have shown that BFCM produces biologically coherent information.

BFCM fixes some of the issues of Trigger. Firstly it does not overestimate the probabilities of regulation as much as Trigger. Trigger only estimates the probability for one of the causal models, whereas BFCM compares a large number of possible models.

5.1 Future Work

In the future we would like to see if the issues with Trigger can be fixed. Currently no underestimates for the proportion of true null hypotheses is available. Thus in the future we'd like to investigate whether or not there are good underestimates of π_0.

We would also like to test BFCM using simulations similar to [21] and [22]. In order to compare BFCM with Trigger on simulated data we would need to use a slightly different approach to simulating the networks. As Trigger uses local false discovery rates in the estimation of the probabilities, it needs to test a large number of hypotheses at the same time. We believe that the comparison between the two algorithms on the yeast dataset provides a sufficient comparison, but a comparison on the simulated data could provide more insight.

Trigger is not the only available method for inferring gene regulation networks. Other methods such as CIT [21] and CMST [22] are also available. We would like to compare BFCM with these methods. As these methods produce p values instead of probability estimates, a framework has to be built which allows us to compare these methods with BFCM.

Bibliography

[1] Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, James D Watson, and AV Grimstone. Molecular biology of the cell (3rd edn). *Trends in Biochemical Sciences*, 20(5):210–210, 1995.

[2] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[3] Nan Bing and Ina Hoeschele. Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics*, 170(2):533–542, 2005.

[4] Elizabeth I Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J Michael Cherry, and Gavin Sherlock. Go::termfinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, 2004.

[5] Rachel B Brem and Leonid Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 102(5):1572–1577, 2005.

[6] Rachel B Brem, Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755, 2002.

[7] Lin S Chen, Frank Emmert-Streib, and John D Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes, additional data file #1: Supplementary text and figures.

[8] Lin S Chen, Frank Emmert-Streib, and John D Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome biology*, 8(10):R219, 2007.

[9] Lin S. Chen, Dipen P. Sangurdekar, and John D. Storey. *trigger: Transcriptional Regulatory Inference from Genetics of Gene ExpRession*. R package version 1.14.0.

[10] Elissa J Chesler, Lu Lu, Siming Shou, Yanhua Qu, Jing Gu, Jintao Wang, Hui Chen Hsu, John D Mountz, Nicole E Baldwin, Michael A Langston, et al. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature genetics*, 37(3):233–242, 2005.

[11] Christophe Cullin, Agnès Baudin-Baillieu, Elisabeth Guillemet, and Odile Ozier-Kalogeropoulos. Functional analysis of ycl09c: evidence for a role as the regulatory subunit of acetolactate synthase. *Yeast*, 12(15):1511–1518, 1996.

[12] Bradley Efron. *Local false discovery rates*. Division of Biostatistics, Stanford University, 2005.

[13] Bradley Efron and Robert Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genetic epidemiology*, 23(1):70–86, 2002.

[14] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160, 2001.

[15] Frank Emmert-Streib, Ricardo de Matos Simoes, Paul Mullan, Benjamin Haibe-Kains, and Matthias Dehmer. The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. *Front Genet*, 5:15, 2014.

[16] Tom Heskes. Causal discovery: the normal case. Technical report, Faculty of Science, Radboud University Nijmegen, February 2014.

[17] David C Kulp and Manjunatha Jagalur. Causal inference of regulator-target pairs by gene mapping of expression phenotypes. *BMC genomics*, 7(1):1, 2006.

[18] Renhua Li, Shirng-Wern Tsaih, Keith Shockley, Ioannis M Stylianou, Jon Wergedal, Beverly Paigen, and Gary A Churchill. Structural model analysis of multiple quantitative traits. *PLoS Genet*, 2(7):e114, 2006.

[19] Gavin MacBeath and Stuart L Schreiber. Printing proteins as microarrays for high-throughput function determination. *Science*, 289(5485):1760–1763, 2000.

[20] Hajime Matsuzaki, Shoulian Dong, Halina Loi, Xiaojun Di, Guoying Liu, Earl Hubbell, Jane Law, Tam Berntsen, Monica Chadha, Henry

Hui, et al. Genotyping over 100,000 snps on a pair of oligonucleotide arrays. *Nature Methods*, 1(2), 2004.

[21] Joshua Millstein, Bin Zhang, Jun Zhu, and Eric E Schadt. Disentangling molecular relationships with a causal inference test. *BMC genetics*, 10(1):23, 2009.

[22] Elias Chaibub Neto, Aimee T Broman, Mark P Keller, Alan D Attie, Bin Zhang, Jun Zhu, and Brian S Yandell. Modeling causality for pairs of phenotypes in system genetics. *Genetics*, 193(3):1003–1013, 2013.

[23] Van Hanh Nguyen and Catherine Matias. On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. *Scandinavian Journal of Statistics*, 41(4):1167–1194, 2014.

[24] Siew Siew Pang and Ronald G Duggleby. Expression, purification, characterization, and reconstitution of the large and small subunits of yeast acetohydroxyacid synthase. *Biochemistry*, 38(16):5222–5231, 1999.

[25] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[26] Stan Pounds and Stephan W Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242, 2003.

[27] Joerg Reinders, Rene P Zahedi, Nikolaus Pfanner, Chris Meisinger, and Albert Sickmann. Toward the complete yeast mitochondrial proteome: multidimensional separation techniques for mitochondrial proteomics. *Journal of proteome research*, 5(7):1543–1554, 2006.

[28] Murray Rosenblatt et al. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

[29] Cosmin Saveanu, Micheline Fromont-Racine, Alexis Harington, Florence Ricard, Abdelkader Namane, and Alain Jacquier. Identification of 12 new yeast mitochondrial ribosomal proteins including 6 that have no prokaryotic homologues. *Journal of Biological Chemistry*, 276(19):15861–15867, 2001.

[30] Eric E Schadt, Stephanie A Monks, Thomas A Drake, Aldons J Lusis, Nam Che, Veronica Colinayo, Thomas G Ruff, Stephen B Milligan, John R Lamb, Guy Cavet, et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302, 2003.

[31] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[32] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[33] Matthew P Wand, James Stephen Marron, and David Ruppert. Transformations in density estimation. *Journal of the American Statistical Association*, 86(414):343–353, 1991.

Appendix A

Appendix

A.1 Statistics

A.1.1 Hypothesis Testing

In statistics we are often concerned with modeling a population or experiment. A statistical model is a pair (S, P) where S is the sample space and P is a set of probability distributions on S. The set P is chosen such that it contains a distribution that approximates the true distribution. Usually the set P is parametrized: $P = \{P_\theta : \theta \in \Theta\}$, where Θ is the set of all possible parameters in the model.

We are interested in inferring more information about the true parameter, θ. One way of doing this is by utilizing a hypothesis test. In a hypothesis test we are concerned with testing a Null Hypothesis, $H_0 : \theta \in \Theta_0$, against an Alternative Hypothesis, $H_1 : \theta \in \Theta_1$. Here $\{\Theta_0, \Theta_1\}$ is a partitioning of the parameter space Θ. Testing can lead to two kinds of conclusions: We either reject H_0 (and accept H_1 as being correct) or we do not reject H_0 (but we do accept H_1 as being incorrect either).

There are two kinds of errors that can be made: A type I error occurs when we reject H_0 even though H_0 is correct, a type II error occurs when we fail to reject H_0 when H_0 is incorrect. A type I error is also called a false discovery.

As the data $X = (X_1, \ldots, X_n)$ can consist of many observations it can be difficult to test the Null Hypothesis. In this case it might useful to summarize (a part of) the data in a single real value $T = T(X)$, which is called a statistic. This statistic can be chosen such that it does not depend on the unknown parameter.

We have not specified yet when the Null Hypothesis should be rejected.

In order to these we a define a rejection region R, if $T \in R$ we reject the Null Hypothesis. We want to choose R such that we minimize the type I and type II errors.

Definition (Power Function)**.** The power function of a test with rejection region R is defined by:

$$\beta(\theta) = P_\theta(T \in R)$$

We want to choose the rejection region R such that it maximizes the power function when $\theta \in \Theta_0$ and minimizes the power function when $\theta \in \Theta_1$. This turns out to be difficult in practice as the power function is usually a continuous, so decreasing the power under the null hypothesis will result in a decrease of the power under alternative hypothesis.

Definition (Size)**.** A test with power function $\beta(\theta)$ is a size α test if:

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$$

We then select a rejection region R which has size α and the maximal power under the alternative hypothesis.

A.1.2 P-values

In the last paragraph we used a rejection region to define a test. The size, α, describes how reliable the test is. If α is low the decision to reject a null hypothesis is convincing, but if α is high the probability of a false positive becomes too high and the rejection becomes unconvincing. Instead of using the size of a test we can also look at the so called p-value.

Definition (Right-tailed p-value)**.** Let t be the observed value of the test statistic T then the right-tailed p-value is defined as

$$p(T) = \sup_{\theta \in \Theta_0} \mathrm{P}_\theta \left(T \geq t\right)$$

If the p-value, p, is low we reject the null hypothesis. It is easy to construct a rejection region with size α using the p-values. If we reject H_0 if and only if $p(T) \leq \alpha$ we get the rejection region $R = \{T : p(T) \leq \alpha\}$. Thus the smaller the p-value is, the stronger the evidence becomes that the null hypothesis is false.

Instead of using a right-tailed p-value we can also use the left-tailed p-value

$$p(T) = \sup_{\theta \in \Theta_0} \mathrm{P}_\theta \left(T \leq t\right)$$

or the two-tailed p-value

$$p(T) = 2\min(\sup_{\theta\in\Theta_0} \mathrm{P}_\theta\left(T \leq t\right), \sup_{\theta\in\Theta_0} \mathrm{P}_\theta\left(T \geq t\right)).$$

A.1.3 Likelihood-Ratio Test

An example of a hypothesis test is the likelihood-ratio test, which is the test used by Trigger. The likelihood-ratio test compares the likelihood of the null hypothesis given the data with the likelihood of the alternative hypothesis given the data. The likelihood-ratio test is defined using the likelihood-ratio statistic:

Definition (Likelihood-Ratio Statistic)**.** Let X be a stochastic vector with density function p_θ, then the likelihood-ratio statistic for the test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is defined as:

$$\Lambda(X) = \frac{\sup_{\theta\in\Theta} L(\theta; X)}{\sup_{\theta\in\Theta_0} L(\theta; X)}$$

Using the asymptotic behavior of two times the log of the likelihood-ratio we can construct a rejection region. Wilks's theorem states that under certain conditions the distribution of $2\log\Lambda(X)$ under the null hypothesis converges to $\chi^2_{k-k_0}$ when $n \to \infty$. Here k is dimension of Θ and k_0 is the dimension of Θ_0. In the case that the null distribution is unknown or cannot be approximated using its asymptotic behaviour, permutation testing can be used to approximate p values. This done by calculating the test statistics on permuted data. These statistics can then be used to calculate the empirical p value.

The likelihood-ratio test is related to the so called maximum likelihood estimators, which maximize the likelihood function with respect to θ.

Definition (Maximum Likelihood Estimator)**.** Let X be a stochastic with density p_θ which depends on $\theta \in \Theta$, the maximum likelihood estimator (MLE) is defined as:

$$T(X) = \arg\sup_{\theta\in\Theta} L(\theta; X)$$

We can now see that likelihood-ratio statistic is simply the likelihood-ratio ratio of the likelihood evaluated at the MLE and the MLE restricted on Θ_0.

We can apply the likelihood ratio test to regression models. Suppose a stochastic variable Y is linearly dependent of the variable X then we can use the following simple linear regression model to model this relationship:

$$Y = \alpha + \beta X + \epsilon$$

where α is called the intercept, β is called the slope and ϵ is the error, which is usually distributed by a normal distribution $N(0, \sigma^2)$. We will only look at the even simpler case in which X is either equal to -1 or equal to 1.

Often we want to test whether Y depends on X, which involves testing whether $\beta \neq 0$. This can be done using a likelihood ratio with $H_0 : \beta = 0$ and $H_1 : \beta \neq 0$. Which gives us the following test statistic:

$$\Lambda = \left(\frac{\sum_{i=1}^{N_{=-1}} (x_{=-1,i} - \overline{x_{=-1}})^2 + \sum_{i=1}^{N_{=1}} (x_{=1,i} - \overline{x_{=1}})^2}{\sum_{i=1}^{N} (x_i - \overline{x})^2} \right)^{\frac{N}{2}}$$

A.2 Test statistics used by Trigger

This appendix gives the test statistics used by Trigger, in order to test whether three models holds.

A.2.1 Notation

We denote the observed transcription levels as t_{ik} for the ith gene and kth sample of the n samples. l_i is equal to -1 if the allele was inherited from the first parent and equal to 1 if it was inherited from the second parent. We use t^a_{ik} with $1 \leq i \leq n_a$ to denote the t_{im} with $l_{im} = -1$ and t^b_{ik} with $1 \leq i \leq n_b$ to denote the t_{im} with $l_{im} = 1$. The average of n samples is denoted as $\overline{t_i}$ and similarly we denote the average of the n_a samples with $l_{ik} = -1$ as $\overline{t^a_i}$.

A.2.2 Local linkage

$$X_i = \left(\frac{\sum_{k=1}^{n} (t_{ik} - \overline{t_i})^2}{\sum_{k=1}^{n_a} (t^a_{ik} - \overline{t_i}^a)^2 + \sum_{k=1}^{n_b} (t^b_{ik} - \overline{t_i}^b)^2} \right)^{\frac{n}{2}}$$

A.2.3 Secondary linkage

$$Y_{ij} = \left(\frac{\hat{\sigma}^2_{j,0} \hat{\sigma}^2_{i,0} - \hat{\sigma}^2_{ij,0}}{\hat{\sigma}^2_j \hat{\sigma}^2_i - \hat{\sigma}^2_{ij}} \right)^{\frac{n}{2}}$$

with:

$$\hat{\sigma}_i^2 = \frac{1}{n}\left(\sum_{k=1}^{n_a}(t_{ik}^a - \overline{t_i^a})^2 + \sum_{k=1}^{n_b}(t_{ik}^b - \overline{t_i^b})^2\right)$$

$$\hat{\sigma}_{ij} = \frac{1}{n}\left(\sum_{k=1}^{n_a}(t_{ik}^a - \overline{t_i^a})(t_{jk}^a - \overline{t_j^a}) + \sum_{k=1}^{n_b}(t_{ik}^b - \overline{t_i^b})(t_{jk}^b - \overline{t_j^b})\right)$$

$$\hat{\sigma}_j^2 = \frac{1}{n}\left(\sum_{k=1}^{n_a}(t_{jk}^a - \overline{t_j^a})^2 + \sum_{k=1}^{n_b}(t_{jk}^b - \overline{t_j^b})^2\right)$$

$$\hat{\sigma}_{i,0}^2 = \frac{1}{n}\left(\sum_{k=1}^{n_a}(t_{ik}^a - \overline{t_i^a})^2 + \sum_{k=1}^{n_b}(t_{ik}^b - \overline{t_i^b})^2\right)$$

$$\hat{\sigma}_{j,0}^2 = \frac{1}{n}\sum_{k=1}^{n}(t_{jk} - \overline{t_j})^2$$

$$\hat{\sigma}_{ij,0} = \frac{1}{n}\left(\sum_{k=1}^{n_a}(t_{ik}^a - \overline{t_i^a})(t_{jk}^a - \overline{t_j}) + \sum_{k=1}^{n_b}(t_{ik}^b - \overline{t_i^b})(t_{jk}^b - \overline{t_j})\right)$$

A.2.4 Conditional independence

$$Z_{ij} = \frac{\frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2)^{\frac{n}{2}}}{(\frac{1}{n_a}\sum_{k=1}^{n_a}(x_k^a - \overline{x^a})^2)^{\frac{n_a}{2}}(\frac{1}{n_b}\sum_{k=1}^{n_b}(x_k^b - \overline{x^b})^2)^{\frac{n_b}{2}}}$$

A.3 Extra Results of CBF

GO term	P value	Cluster frequency	Background frequency	FDR (%)	Genes
structural constituent of ribosome	2.8751e-13	14 out of 26	233 out of 7166	0.00%	NAM9, MRPS8, RSM18, RSM28, MRP10, MNP1, MRPL3, MRPS5, MRP49, MRPL27, MRPS17, MRP1, RSM7, MRPL11
structural molecule activity	1.609e-10	14 out of 26	368 out of 7166	0.00%	NAM9, MRPS8, RSM18, RSM28, MRP10, MNP1, MRPL3, MRPS5, MRP49, MRPL27, MRPS17, MRP1, RSM7, MRPL11

Table A.1: GO Terms for biological functions for NAM9 and top 25 regulated genes found by CBF

GO term	P value	Cluster frequency	Background frequency	FDR (%)	Genes
organellar ribosome	3.209e-21	15 out of 26	91 out of 7166	0.00%	CBP6, NAM9, MRPS8, RSM18, RSM28, MRP10, MRPL3, MNP1, MRPS5, MRP49, MRPL27, MRPS17, MRP1, RSM7, MRPL11
mitochondrial ribosome	3.209e-21	15 out of 26	91 out of 7166	0.00%	CBP6, NAM9, MRPS8, RSM18, RSM28, MRP10, MRPL3, MNP1, MRPS5, MRP49, MRPL27, MRPS17, MRP1, RSM7, MRPL11
mitochondrial matrix	1.875e-20	18 out of 26	222 out of 7166	0.00%	CBP6, RSM18, RSM28, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MNP1, MRPL3, MRP49, MRPL27, RSM7, MRPL11
mitochondrial part	1.212e-18	22 out of 26	617 out of 7166	0.00%	CBP6, UBP16, RSM18, RSM28, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MDM35, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
mitochondrion	2.721e-17	25 out of 26	1195 out of 7166	0.00%	CBP6, UBP16, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MTO1, MDM35, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
organellar small ribosomal subunit	4.889e-14	9 out of 26	34 out of 7166	0.00%	NAM9, MRPS8, RSM18, RSM28, MRP10, MRPS5, MRP1, RSM7, MRPS17
mitochondrial small ribosomal subunit	4.889e-14	9 out of 26	34 out of 7166	0.00%	NAM9, MRPS8, RSM18, RSM28, MRP10, MRPS5, MRP1, RSM7, MRPS17
ribosomal subunit	9.849e-13	14 out of 26	254 out of 7166	0.00%	NAM9, MRPS8, RSM18, RSM28, MRP10, MNP1, MRPL3, MRPS5, MRP49, MRPL27, MRPS17, MRP1, RSM7, MRPL11
ribosome	5.361e-12	15 out of 26	362 out of 7166	0.00%	CBP6, NAM9, MRPS8, RSM18, RSM28, MRP10, MRPL3, MNP1, MRPS5, MRP49, MRPL27, MRPS17, MRP1, RSM7, MRPL11
membrane-enclosed lumen	1.112e-9	20 out of 26	1211 out of 7166	0.00%	CBP6, RSM18, RSM28, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MDM35, YNR020C, MNP1, MRPL3, MRP49, MRPL27, RSM7, MRPL11
small ribosomal subunit	1.538e-9	9 out of 26	100 out of 7166	0.00%	NAM9, MRPS8, RSM18, RSM28, MRP10, MRPS5, MRP1, RSM7, MRPS17
organelle lumen	8.654e-8	18 out of 26	1154 out of 7166	0.00%	CBP6, RSM18, RSM28, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MNP1, MRPL3, MRP49, MRPL27, RSM7, MRPL11
intracellular organelle lumen	8.654e-8	18 out of 26	1154 out of 7166	0.00%	CBP6, RSM18, RSM28, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MNP1, MRPL3, MRP49, MRPL27, RSM7, MRPL11
intracellular ribonucleoprotein complex	4.675e-7	15 out of 26	800 out of 7166	0.00%	CBP6, NAM9, MRPS8, RSM18, RSM28, MRP10, MRPL3, MNP1, MRPS5, MRP49, MRPL27, MRPS17, MRP1, RSM7, MRPL11
ribonucleoprotein complex	4.675e-7	15 out of 26	800 out of 7166	0.00%	CBP6, NAM9, MRPS8, RSM18, RSM28, MRP10, MRPL3, MNP1, MRPS5, MRP49, MRPL27, MRPS17, MRP1, RSM7, MRPL11
cytoplasmic part	1.711e-6	25 out of 26	3262 out of 7166	0.00%	CBP6, UBP16, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MTO1, MDM35, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
intracellular organelle part	0.00055	22 out of 26	3055 out of 7166	0.00%	CBP6, UBP16, RSM18, RSM28, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MDM35, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
organelle part	0.00060	22 out of 26	3060 out of 7166	0.00%	CBP6, UBP16, RSM18, RSM28, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MDM35, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
intracellular membrane-bounded organelle	0.00083	25 out of 26	4216 out of 7166	0.00%	CBP6, UBP16, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MTO1, MDM35, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
membrane-bounded organelle	0.00088	25 out of 26	4228 out of 7166	0.00%	CBP6, UBP16, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MTO1, MDM35, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
non-membrane-bounded organelle	0.00096	15 out of 26	1420 out of 7166	0.00%	CBP6, NAM9, MRPS8, RSM18, RSM28, MRP10, MRPL3, MNP1, MRPS5, MRP49, MRPL27, MRPS17, MRP1, RSM7, MRPL11
intracellular non-membrane-bounded organelle	0.00096	15 out of 26	1420 out of 7166	0.00%	CBP6, NAM9, MRPS8, RSM18, RSM28, MRP10, MRPL3, MNP1, MRPS5, MRP49, MRPL27, MRPS17, MRP1, RSM7, MRPL11
cytoplasm	0.0033	25 out of 26	4471 out of 7166	0.00%	CBP6, UBP16, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MTO1, MDM35, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
intracellular organelle	0.00512	25 out of 26	4555 out of 7166	0.00%	CBP6, UBP16, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MTO1, MDM35, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
organelle	0.00541	25 out of 26	4566 out of 7166	0.07%	CBP6, UBP16, RSM18, MEF1, RSM28, IFM1, MRP10, MRPS5, DLD2, MRPS17, MRP1, DIA4, NAM9, MRPS8, ACN9, MTO1, MDM35, MNP1, YNR020C, MRPL3, MRP49, MRPL27, RSM7, COX10, MRPL11
large ribosomal subunit	0.00843	5 out of 26	154 out of 7166	0.07%	MRPL3, MNP1, MRP49, MRPL27, MRPL11

Table A.2: GO Terms for biological components for NAM9 and top 25 regulated genes found by CBF

GO term	P value	Cluster frequency	Background frequency	FDR (%)	Genes
alpha-amino acid biosynthetic process	6.412e-11	11 out of 26	135 out of 7166	0.00%	ASN1, HIS4, THR1, HOM3, ARG1, HIS1, ILV6, LYS21, CPA2, ARG2, TRP4
cellular amino acid biosynthetic process	1.316e-10	11 out of 26	144 out of 7166	0.00%	ASN1, HIS4, THR1, HOM3, ARG1, HIS1, ILV6, LYS21, CPA2, ARG2, TRP4
alpha-amino acid metabolic process	1.558e-10	12 out of 26	199 out of 7166	0.00%	ASN1, HIS4, THR1, HOM3, LAP3, ARG1, HIS1, ILV6, LYS21, CPA2, ARG2, TRP4
cellular amino acid metabolic process	4.668e-9	12 out of 26	265 out of 7166	0.00%	ASN1, HIS4, THR1, HOM3, LAP3, ARG1, HIS1, ILV6, LYS21, CPA2, ARG2, TRP4
organic acid biosynthetic process	6.4000e-9	11 out of 26	205 out of 7166	0.00%	ASN1, HIS4, THR1, HOM3, ARG1, HIS1, ILV6, LYS21, CPA2, ARG2, TRP4
carboxylic acid biosynthetic process	6.4000e-9	11 out of 26	205 out of 7166	0.00%	ASN1, HIS4, THR1, HOM3, ARG1, HIS1, ILV6, LYS21, CPA2, ARG2, TRP4
small molecule biosynthetic process	1.452e-7	12 out of 26	356 out of 7166	0.00%	ASN1, HIS4, THR1, HOM3, ARG1, HIS1, ILV6, LYS21, RIB3, CPA2, ARG2, TRP4
carboxylic acid metabolic process	1.631e-6	12 out of 26	440 out of 7166	0.00%	ASN1, HIS4, THR1, HOM3, LAP3, ARG1, HIS1, ILV6, LYS21, CPA2, ARG2, TRP4
oxoacid metabolic process	2.442e-6	12 out of 26	456 out of 7166	0.00%	ASN1, HIS4, THR1, HOM3, LAP3, ARG1, HIS1, ILV6, LYS21, CPA2, ARG2, TRP4
organic acid metabolic process	2.565e-6	12 out of 26	458 out of 7166	0.00%	ASN1, HIS4, THR1, HOM3, LAP3, ARG1, HIS1, ILV6, LYS21, CPA2, ARG2, TRP4
small molecule metabolic process	3.045e-6	15 out of 26	837 out of 7166	0.00%	ASN1, HIS4, THR1, HOM3, LAP3, ARG1, HIS1, ILV6, LYS21, RIB3, URA10, CPA2, ZWF1, ARG2, TRP4
single-organism biosynthetic process	0.0008735	13 out of 26	934 out of 7166	0.00%	ASN1, HIS4, THR1, HOM3, ARG1, HIS1, ILV6, LYS21, RIB3, URA10, CPA2, ARG2, TRP4
organonitrogen compound metabolic process	0.0021103	15 out of 26	1369 out of 7166	0.00%	ASN1, HIS4, THR1, HOM3, LAP3, ARG1, HIS1, ILV6, LYS21, RIB3, URA10, CPA2, ZWF1, ARG2, TRP4
organonitrogen compound biosynthetic process	0.008190	13 out of 26	1143 out of 7166	0.00%	ASN1, HIS4, THR1, HOM3, ARG1, HIS1, ILV6, LYS21, RIB3, URA10, CPA2, ARG2, TRP4

Table A.3: GO Terms for biological processes for ILV6 and top 25 regulated genes found by CBF

GO term	P value	Cluster frequency	Background frequency	FDR (%)	Genes
amino acid binding	0.004223	3 out of 26	23 out of 7166	0.00%	HOM3, ARG2, ILV6
carboxylic acid binding	0.006155	3 out of 26	26 out of 7166	1.00%	HOM3, ARG2, ILV6
organic acid binding	0.006155	3 out of 26	26 out of 7166	0.80%	HOM3, ARG2, ILV6

Table A.4: GO Terms for biological functions for ILV6 and top 25 regulated genes found by CBF