Detecting Active Pathways In Gene Sets

Matthew T. Lomont

April 30, 2010

1 Introduction

1.1 Gene Expression Levels

Biological databases can be used to define groups of related genes that will be referred to as gene sets, or pathways. We can study individual gene expression levels to determine whether a pathway contains active genes, and what exactly these genes are[1]. But there are a few obstacles that arise when trying to determine exactly which genes are active. Biological processes in cells often affect sets of genes in unison[2], and conducting analysis on a single gene-by-gene basis does not account for any sort of correlation between genes. A simple case could be that the rise in the expression level of one gene leads to a rise/decline in the expression of another. But it is also likely that much more complicated correlation structures exist within pathways. When analyzing gene sets, it is likely we will obtain different results if the data are correlated as opposed to being independent. It is important that any method used is robust to different correlation structures.

A second obstacle is the difficulty that arises with multiple hypothesis testing. Often microarray data consists of a large number of genes, but only a small sample size of expression levels for each gene[1]. This creates a relatively big potential for falsely discovering genes to be significant that are not, in fact, significant. A typical approach to regulate this would be to simply pre-determine the type I error rate (or a significance level) to an acceptable level, say .05. This seems logical on a per-comparison basis, but issues arise when large-scale multiple hypothesis testing uses this approach. For instance, take a gene set with 1000 genes, none of which are actually significant. Simply by chance we would expect to falsely discover 50 significant genes.

Another common approach would be to correct the type I error rate for multiple hypothesis testing. A good example is the Bonferroni correction, which would effectively use a per comparison significance level of our type I error rate divided by the number of tests. In the previous example, our new per comparison significance level would be .00005. This method is far too conservative, and has little power in rejecting a false null hypothesis[3]. The type I error rate is controlled well, but it is controlled at the expense of the type II error rate (false negative rate). A proper test for this sort of data must focus on regulating the type I error rate while maintaining high power in order to detect significant genes.

1.3 False Discovery Rate

Table 1 gives some insight into the multiple-hypothesis testing problem

Table 1: Possible Results From 'n' Hypothesis tests

	Failed to Reject H_0	Reject H_0	Total
H_0 True	True Negatives (TN)	False Positives (FP)	n_0
H_0 False	False Negatives (FN)	True Positives (TP)	n_1
	Negatives (N)	Positives (P)	n

Using the uncorrected, per comparison, type I error threshold described above, the false positive rate is regulated such that:

$$\boldsymbol{P}(\ \mathrm{FP}_i > 0\ ) \leq \alpha$$

$$1 \leq i \leq \mathrm{n}.$$

More simply put, each FP_i is a binary variable, either equal to zero, or one (it was either a false positive or not). The individual probability that it was a false positive is less than α for each FP_i. Using the corrected, per comparison, type I error threshold, the false positive rate is regulated such that:

$$\alpha^* = \alpha / n$$

$$\boldsymbol{P}(\ FP_i > 0\ ) \leq \alpha^*$$

$$1 \leq i \leq n.$$

This is a much more conservative method when the number of hypotheses tests, n, becomes large.

Another proposed method for regulating false discoveries is to control the false discovery rate (FDR). Benjamini and Hochberg[4] were among the first to implement the FDR method, and described an approach such that:

$$\boldsymbol{E}[\ FP / P\ ] \leq \alpha.$$

Simply put, this method requires that the expected proportion of false discoveries is less than some threshold, α. While there are a variety of different ways to implement an FDR procedure, we will only focus on an FDR controlling algorithm based on that used in Benjamini & Hochberg (1995)[4]. After retrieving a set of P-values from n hypothesis tests, rank them $p_1 \leq .. \ p_i \ .. \leq p_n$. A check value, C_i , is then calculated for each p_i:

$$C_i = q * i / n.$$

where q is some set significance level for the FDR procedure. We then compute:

$$K = \max\{1 \leq k \leq n : p_i \leq C_i\}.$$

and reject the null hypothesis for the first K genes in the ordered set of P-values. This method is robust in that it maintains high power, while ensuring a low false positive rate[3].

2 Exploration of Enrichment Scores

2.1 The Maxmean

One area of interest was simply to determine whether or not a pathway contains active genes. Many approaches begin by computing a two-sample t-statistic, z_i, for each gene. The z-values can then be used to form some enrichment score for each pathway. The enrichment score is then compared against some predefined cutoff to determine whether the pathway is active (significant)[1]. In Gene Set Analysis (GSA), Efron and Tibshirani[1] use an approach they call the *maxmean*. In a given gene set, they first compute z_i for each gene. The maxmean then sums the positive components of the z_i and the negative components of the z_i. Then the positive and negative sums are "averaged" by dividing them by the *entire* number of genes in the gene set. The maxmean enrichment score is then set equal to the maximum of the absolute value of these positive and negative averages. The exact formula takes the form:

$$\max\left\{ \frac{\sum_{i=0}^{k} z_i}{k} : z_i > 0, \frac{\sum_{i=0}^{k} |z_i|}{k} : z_i < 0 \right\}.$$

For example, If there were 20 genes, with 1 score of -1.5, and 19 scores of .1, the score would be equal to 19(.1)/20 = .095 since it is larger than 1(1.5)/20 = .075. The authors claim that this approach is robust by design in that it does not allow a few large positive, or negative, scores to dominate. For a much more detailed description of the maxmean approach refer to Efron and Tibshirani (2007) pages 3-6, and 16-17[1].

2.2 Proposed Enrichment Scores

We decided to test the maxmean method against a few ideas of our own. After testing a variety of approaches, we decided to focus on two enrichment scores that could contend with the maxmean. The first method was simply to use the square of the z_i to see how it compared with the maxmean, which uses the un-squared Z-values. By squaring the Z-values we hoped to inflate the differences between the treatment and control groups, making it easier to detect significant pathways.

Our second approach was a little more complicated. We wanted to compare the maxmean method against the P-values corresponding to the Z-values. We transformed the P-values with the intent to inflate the differences between the treatment and control groups. We tried several transformations, and settled on one method that worked best. We first calculated the two-sided P-value for each z_i statistic, then adjusted the P-value using the following method:

$$p_i = 2\Phi(-|z_i|)$$

$$e^{\sqrt{|\log(p_i)|}}.$$

However, to find comparable methods to the maxmean, we had to implement a "maxmean twist" on each of our methods. We did this by setting our enrichment scores equal to:

$$\max\left\{\frac{\sum_{i=0}^{k} z_i^2}{k} : z_i > 0, \frac{\sum_{i=0}^{k} z_i^2}{k} : z_i < 0\right\}$$

$$\max\left\{\frac{\sum_{i=0}^{k} e^{\sqrt{|\log(p_i)|}}}{k} : z_i > 0, \frac{\sum_{i=0}^{k} e^{\sqrt{|\log(p_i)|}}}{k} : z_i < 0\right\}.$$

2.3 Simulation Setup

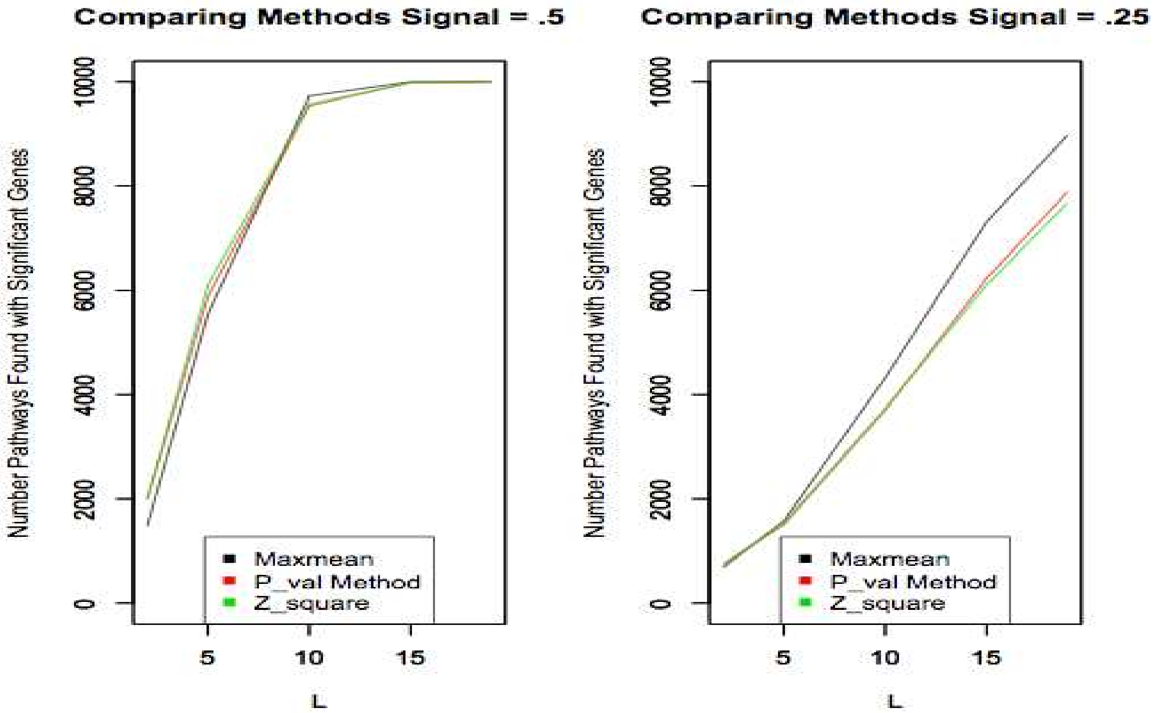
To set up our experiment we wanted to have gene expression levels for 20 genes and 40 samples for each gene. We generated this data from an i.i.d. normal sample, N(0,1). We set up our data in a matrix such that there were 20 rows, one for each gene, and 40 columns, one for each sample of each gene. The first 20 columns were considered to be our treatment group, with the second twenty columns considered to be our control group. In order to simulate a certain set of genes, L, to be "active" genes, we added a treatment (or signal) to the first L genes in the treatment group. We varied the number of active genes, L, between 2, 5, 10, 15, and 19. We also varied the signal added to the active genes between .1, .25, .5, 1, and 1.5. In any given scenario, we can then calculate a t-statistic, z_i, for each of the 20 genes. We replicated this process 10,000 times to ensure accurate results.

For each enrichment score method, we developed a predetermined cutoff for significance based at the 95% confidence level. We can form the null distribution of each enrichment score by adding a treatment of zero. Since there were 10,000 replications, we generated the data with a treatment of zero and order-ranked the test statistics. By chance (at the 95% confidence level) we would expect about 500 of the 10,000 pathways to be falsely identified as having significant genes. For each of our methods, we used the 9,500 largest statistic as our cutoff value to determine what we would declare significant once we added a treatment across a specific number of genes.

2.4 Results

When we add the signal to a large number of genes in the treatment group, the maxmean method always seems to work better than our suggested methods. When the signal is large (1.5), all methods seems to work well in consistently determining all 10,000 pathways to contain significant genes. However, our methods appear to become comparable when we use a signal around .5 and number of treated genes (L) is small. The results if we hold the signal constant at .5, and .25, while varying L, can be seen in Figure 1.

Figure 1: Comparing Methods that Determine Significant pathways.



We can see that our methods do slightly better with a signal of .5, but certainly not enough to claim that they are better than the maxmean. When using a signal of .25, our methods become comparable for L equal to 2 or 5, but the maxmean does far better as L grows. While our methods were comparable to the maxmean under some circumstances, we were not able to find a method that consistently did significantly better in any of the scenarios varying the signal and value of L.

3 Detecting Active Sub Pathways

3.1 Our Adjusted P-Value

Our early exploration only focused on determining whether or not a gene set contained active genes. The larger area of interest is making sure we can consistently detect all of the actually active genes, while not falsely detecting inactive genes. FDR approaches mainly focus on low false discovery, but this may lead to too conservative estimates of what actually is significant. We decided to focus on a method that adjusted the t-statistic, effectively adjusting the P-values. We calculated our z_i according to:

$$Z_i^\alpha = \frac{\left[\lambda(1-\lambda)n\right]^{\left(\frac{1}{2}-\alpha\right)}\left[\overline{Y_i}-\overline{X_i}\right]}{\hat{\sigma}_{pooled}}.$$

where $\lambda = .5$, Y is the treatment group, X is the control group, and n is the total number of samples (twenty in the treatment group, twenty in the control group). The pooled variance equals

$$\hat{\sigma}^2_{pooled} = \frac{1}{n-2}\left[\sum_{j=1}^{\lambda n}\left(X_{ij}-\overline{X_i}\right)^2 + \sum_{j=1}^{(1-\lambda)n}\left(Y_{ij}-\overline{Y_i}\right)^2\right].$$

When α is set to zero in this equation, the first equation gives the regular t-statistic. When α is varied ($0 \le \alpha \le .5$) the t-statistic gets slightly adjusted. We focused on varying α between 1/12, 1/6, 1/4, and 1/3 to test different adjustments. From the t-statistic we determine our one-sided P-value using:

$$P_i^\alpha = 1 - \Phi(Z_i^\alpha)$$

3.2 Simulation Setup and Measures

We set up our sample data in the exact same way as described before, generating data on 20 genes and 40 samples using i.i.d. normal data with mean zero and variance one. We used the same treatments (signals) and varied the number of genes treated (L) using the same values. We then calculated the z_i, using our new adjusted method, and used these to calculate our adjusted P-values.

We were also interested in testing our methods on correlated data. We implemented a simple correlation structure where Z_1, Z_2, … Z_{20} is distributed such that:

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \mathrm{M} \\ Z_{20} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \mathrm{M} \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \rho & \ldots & \mathrm{O} & \rho \\ \rho & \ldots & \rho & 1 \end{bmatrix} \right).$$

This will illustrates twenty gene expressions for one individual with correlation ρ. To do this, we first generated $U_1, U_2, \ldots U_{20} \sim$ i.i.d. N(0,1) and a single Z ~ N(0,1) . After setting ρ, the desired correlation, we generated our sample data according to:

$$\sqrt{1-\rho} \begin{pmatrix} U_1 \\ U_2 \\ \mathrm{M} \\ U_{20} \end{pmatrix} + \sqrt{\rho} \begin{pmatrix} Z \\ Z \\ \mathrm{M} \\ Z \end{pmatrix} = \begin{pmatrix} Z_1 \\ Z_2 \\ \mathrm{M} \\ Z_{20} \end{pmatrix}.$$

$Z_{1,} \ldots, Z_{20}$ represent 20 sample expression levels for one individual. We repeated this 20 times to achieve 20 sample expression levels for 20 individuals to make up our control data. This process was then repeated to obtain 20 samples of 20 individual for our treatment data. Once again, a signal was added to the first L genes of the treatment data to simulate active genes. This whole procedure was replicated 10,000 times.

Consider active genes to form some group, C. Then inactive genes can be grouped into C^C, the complement of C, such that $| C \cup C^C |= 20$, the total number of genes in the sample. For genes in C (active), we expect low P-values close to zero. P-values for genes in C^C (inactive) are expected to converge to a value of .5. Ideally, we would like

$$\sum_{i \in C} \left(P_i^\alpha - 0\right)^2 + \sum_{i \in C^C} \left(P_i^\alpha - .5\right)^2.$$

to equal zero. We want to choose our estimated C, $\hat{C}$, in order to minimize the equation. For our purposes, minimizing the above equation is the same as minimizing:

$$\sum_{i \in C} \left(P_i^\alpha - .25\right).$$

This equation is minimized when we only include genes with adjusted P-values less than or equal to .25 in our $\hat{C}$.

In a real world example, we are unsure of how many, and which genes, in a gene set are actually active. With our simulated data, we do know this. It is simple to determine which genes in $\hat{C}$ are correctly chosen, and which are not. To measure the performance of methods estimating $\hat{C}$ we decided to calculate F_1 scores. An F_1 score is computed using the *precision* and *recall* of a test. The precision is computed by the number of correct results divided by the total number of significant results returned. The recall is computed by the number of correct results divided by the total number of results that are actually significant. The actual F_1 score is then computed using the harmonic mean of precision, p, and recall, r.

$$p = \frac{|\hat{C} \cap C|}{|\hat{C}|} \quad , r = \frac{|\hat{C} \cap C|}{|C|} \quad , \quad F = 2 * \frac{p * r}{p + r}.$$

A precision of one means that all genes in $\hat{C}$ (discoveries) are in fact true discoveries, but says nothing about the number of false negatives. A recall of one means that all active genes were discovered, but says nothing about the number of false discoveries. The F_1 score takes into account both type I and type II errors, and F_1 scores close to one indicate that the method is accurate in keeping both of these low.

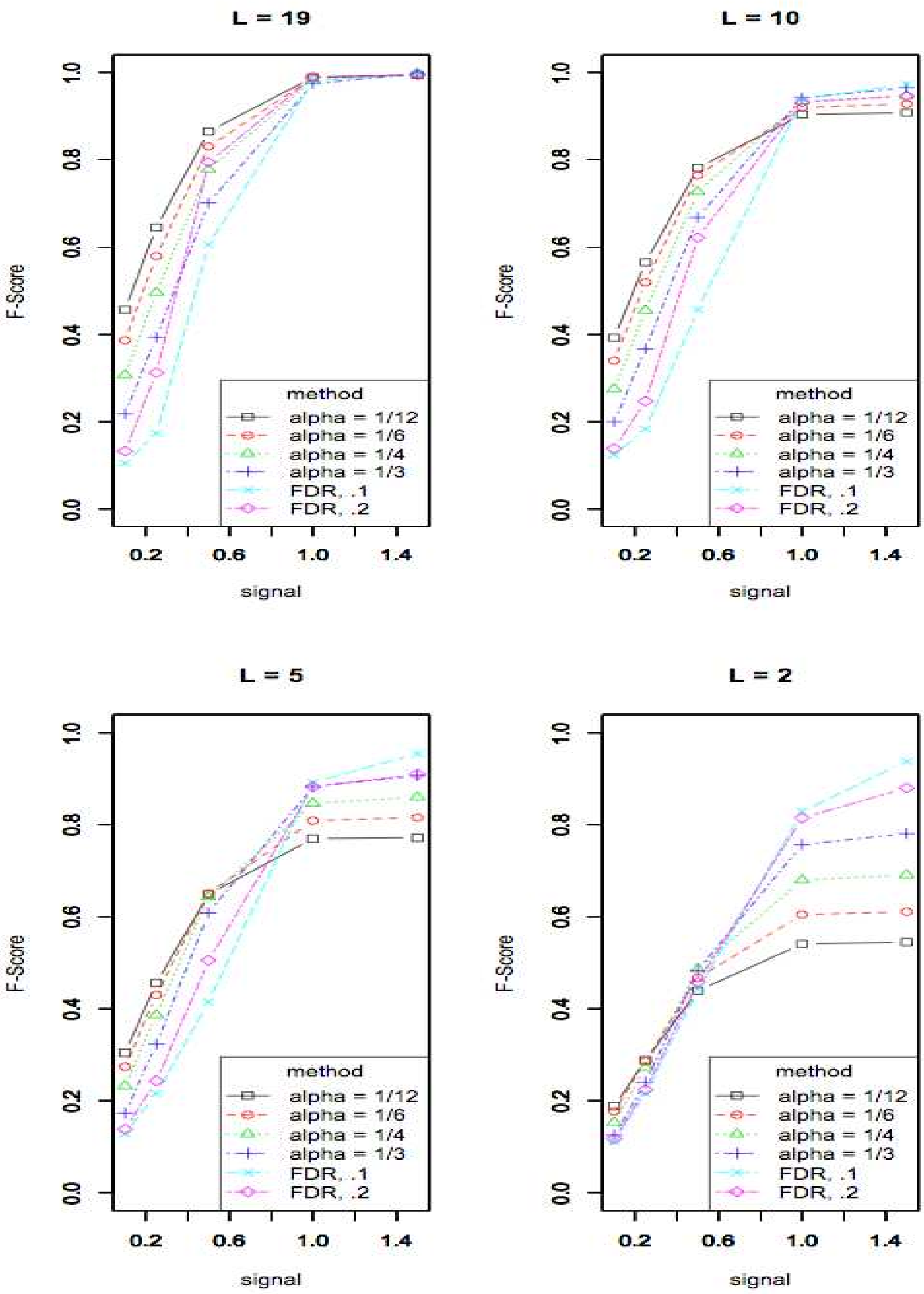
3.3 Determining α

We first need to determine which α is most effective at obtaining high F_1 scores. I will consider a scenario to be a specific signal being added to a specific number of genes (L). We tested five different signals (1.5, 1, .5, .25, .1) and five different values for L (19, 15, 10, 5, 2) giving us 25 scenarios to test for each method. The two main things we are concerned with are the F_1 score, and the estimated $\hat{C}$. By replicating each scenario 10,000 times and taking the mean for each of these values we expect to obtain dependable results.

We begin by observing the results from uncorrelated data (equivalent to $\rho = 0$). Figure 2 shows the results of a few scenarios to get a better idea of how our different α's perform with regards to the F_1 score (for now, disregard FDR).

Figure 2: Comparing Values of α

Uncorrelated Data

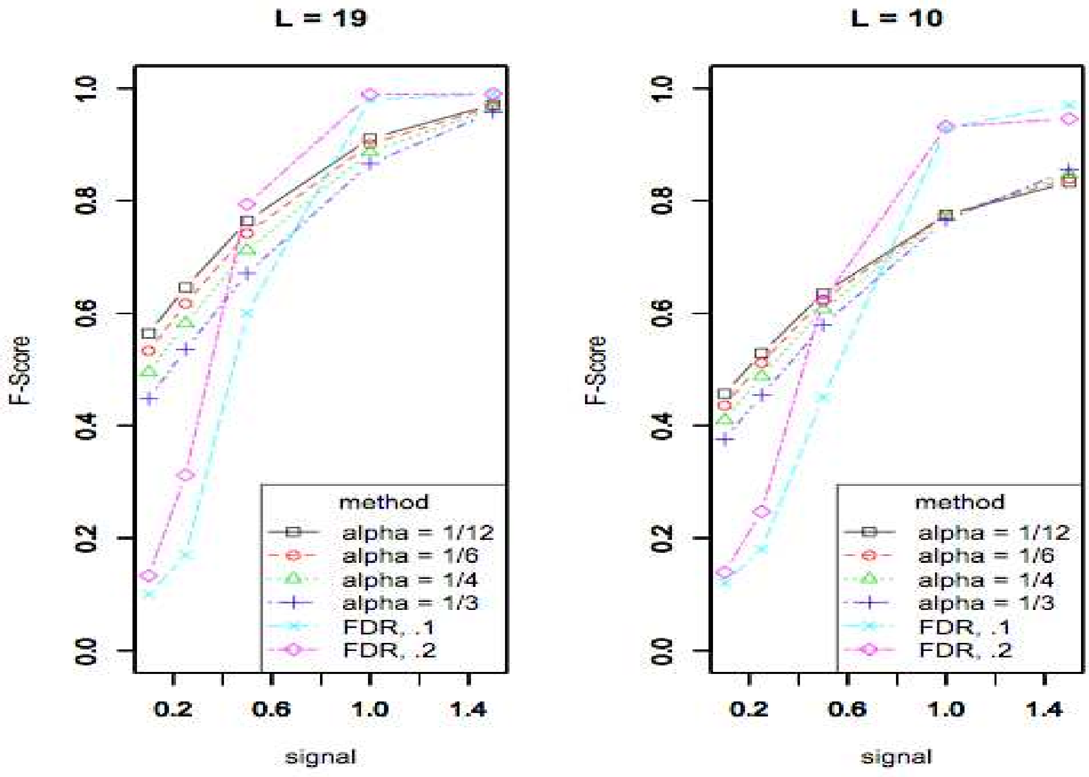


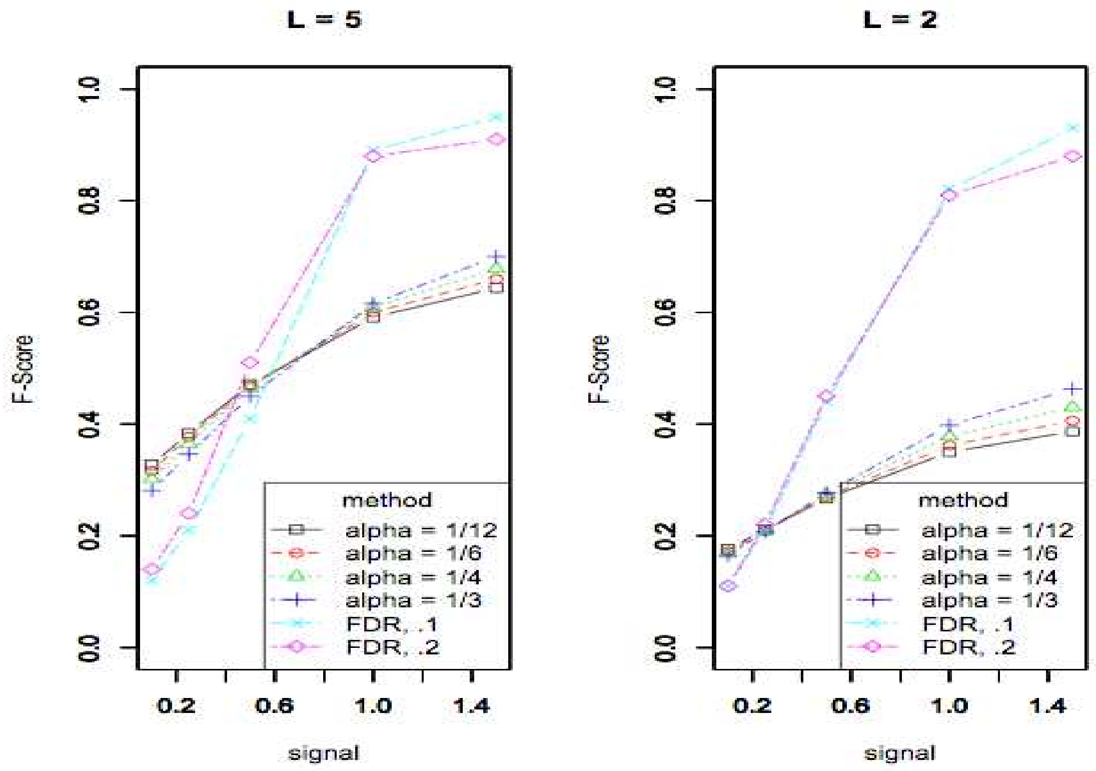
The first thing to note is that when L is high, and the signal is high, all values of α give comparable results. Also, for small values of L and smaller signals, none of the α's stand out distinctly from the rest. The biggest differences occur when L is high and the signal is low, and when L is low and the signal is high. When L is high and the signal is low, α = 1/12 gives the highest F_1 values. When L is low and the signal is high, α = 1/3 gives distinctly higher F_1 values than the others. No values of α stand out as being the best across the board from these results.

Next we focus on the correlated data, generated from the simple structure previously described. We can compare across methods, at set correlations, to determine which α performs best. Figure 3 compares the performance of different values of α holding correlation constant at .2.

Figure 3: Comparing Values of α

Correlation = .2

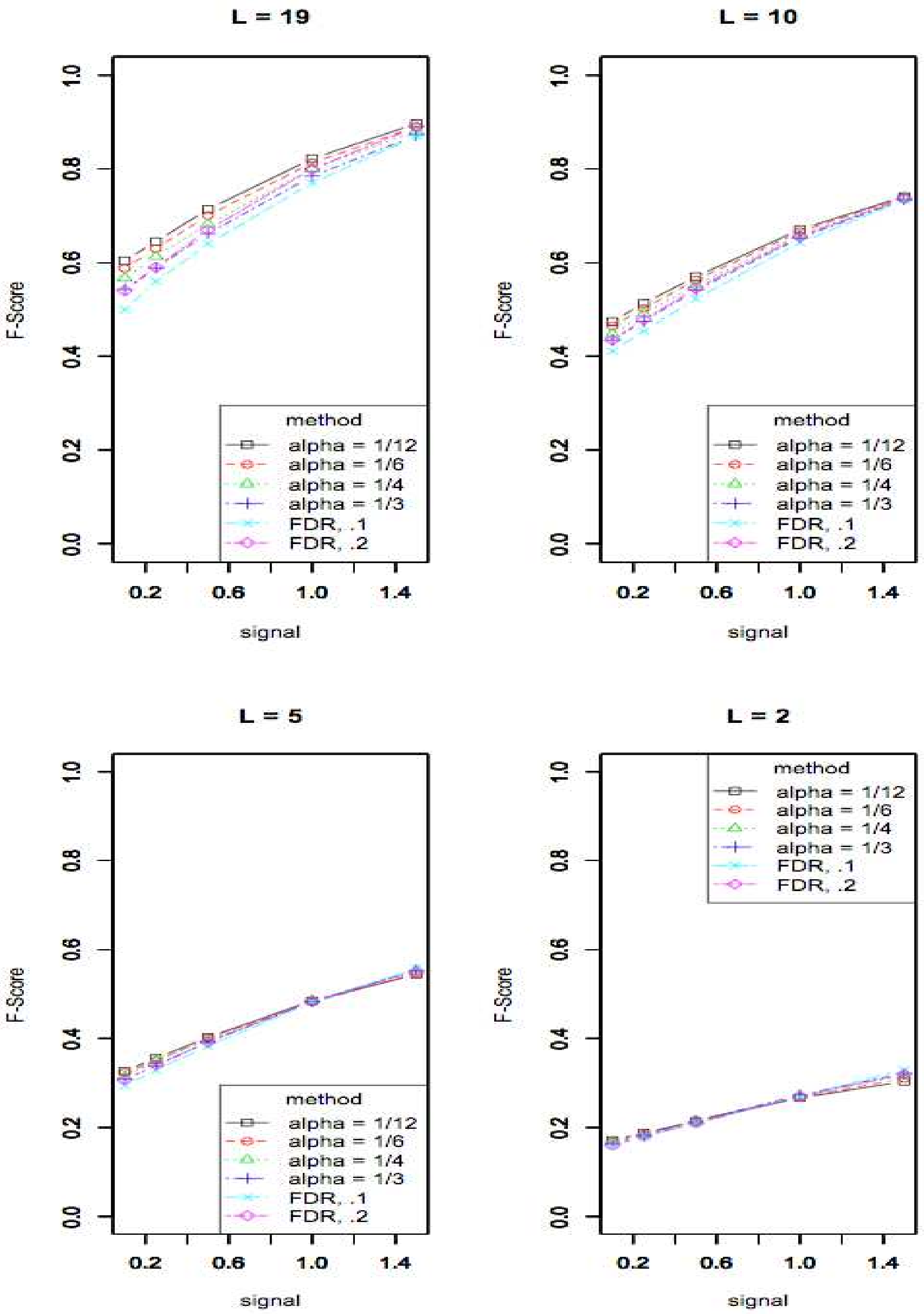




It is clear that the same trends still hold from the uncorrelated data, but the F scores are becoming much more similar for different values of α. The biggest differences are still found when L is high and the signal is low (α = 1/12 being the best), as well as when L is low and the signal is high (α = 1/3 being the best). It is also clear that F scores at higher signals have fallen across all values of L, while F scores at low signals seem to be less effected. To see if these trends hold, we tried a higher correlation. Figure 4 compares the performance of our α's using a correlation of .8.

Figure 4: Comparing Values of α

Correlation = .8

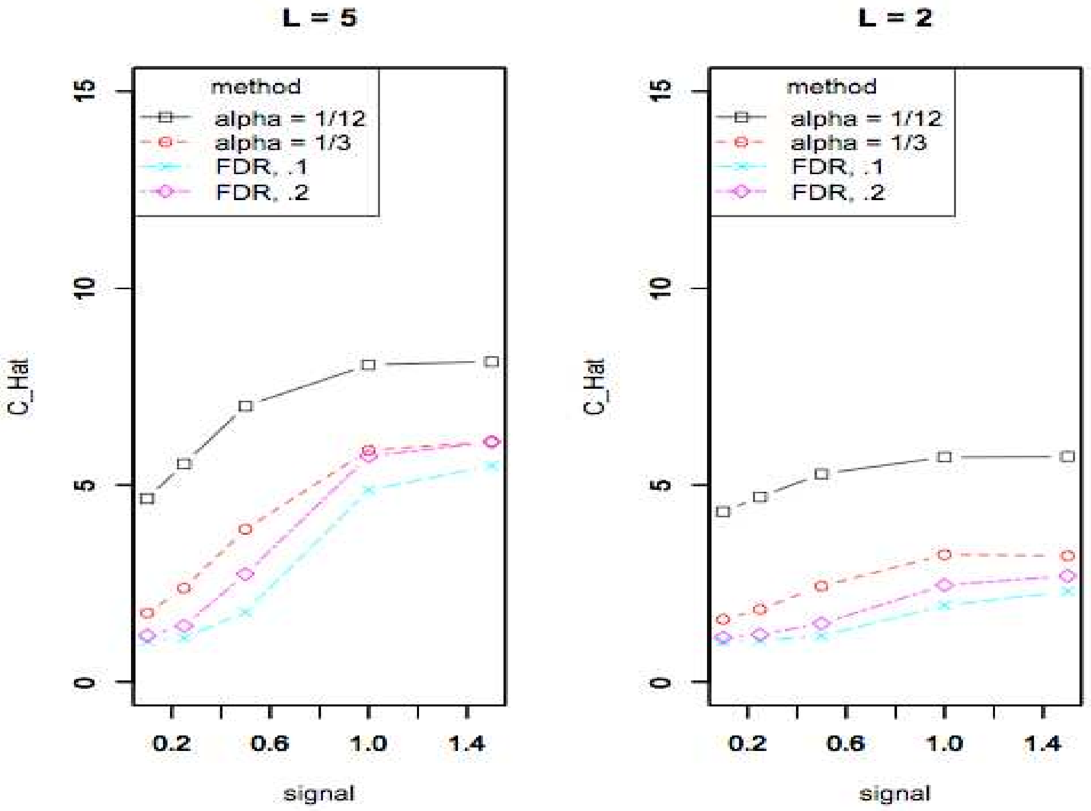


These results indicate that there is little to no difference in the effectiveness of the different α's. Another interesting thing to note is that our F scores for low signals increase slightly with correlated data, while F scores for high signals drop more rapidly. According to these results, as data becomes more highly correlated the value of α seems to matter less, as all values report similar F scores.

When we look at the reported magnitude of $\hat{C}$ (the number of genes we declare active) we find a different story. We would like the magnitude of $\hat{C}$ to be close to L, the actual size of the active set. For simplicity, we can focus on α = 1/3 and α = 1/12. Figure 5 shows the results for L = 5, and L = 2 (For now, disregard the FDR methods).

Figure 5: Estimating |$\hat{C}$|

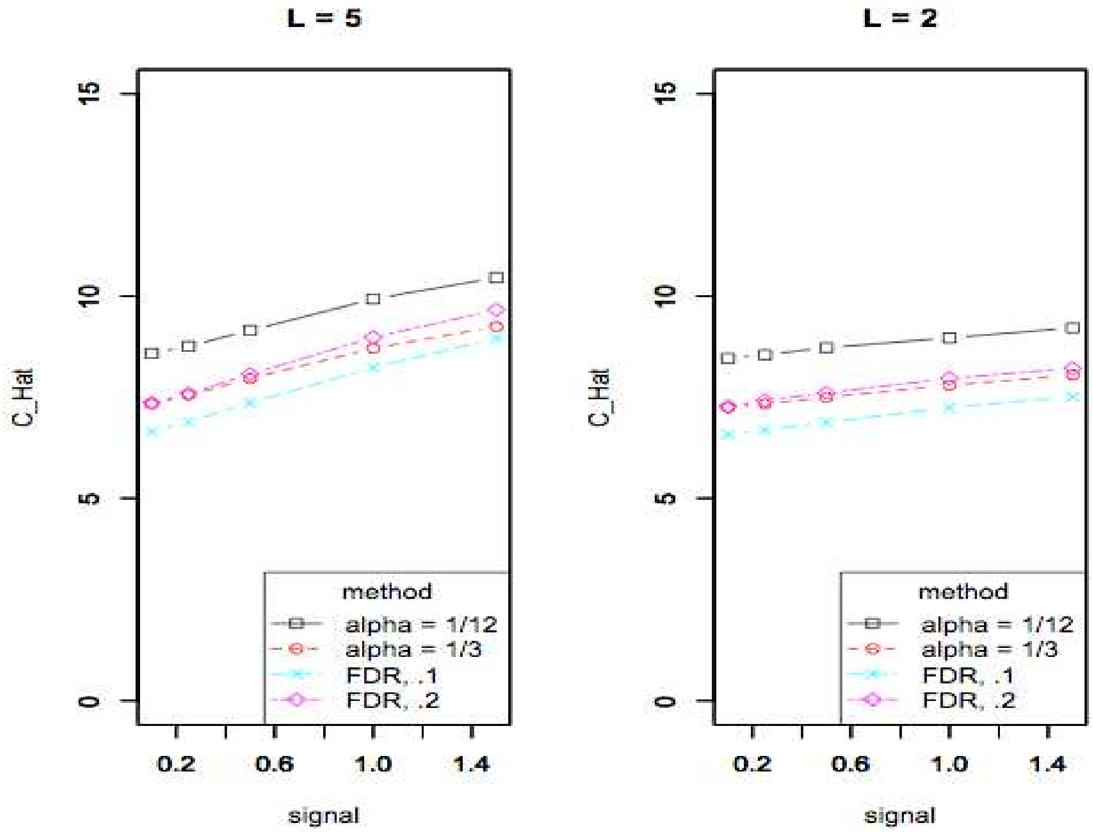
Uncorrelated Data



This suggests that α = 1/3 tends to be more conservative in its estimate of the size of the active pathway. At L = 2, there are only two significant genes, but α = 1/12 consistently estimates the magnitude of $\hat{C}$ between 4 and 5, while α = 1/3 always estimates around two. For high signals and L = 5, α = 1/3 estimates right around 5, while α = 1/12 is once again over-estimating the active set. We can check to see if this pattern holds for the correlated data. Figure 6 shows the estimates of $\hat{C}$ using highly correlated data.

Figure 6: Estimating $|\hat{C}|$

Correlation = .8

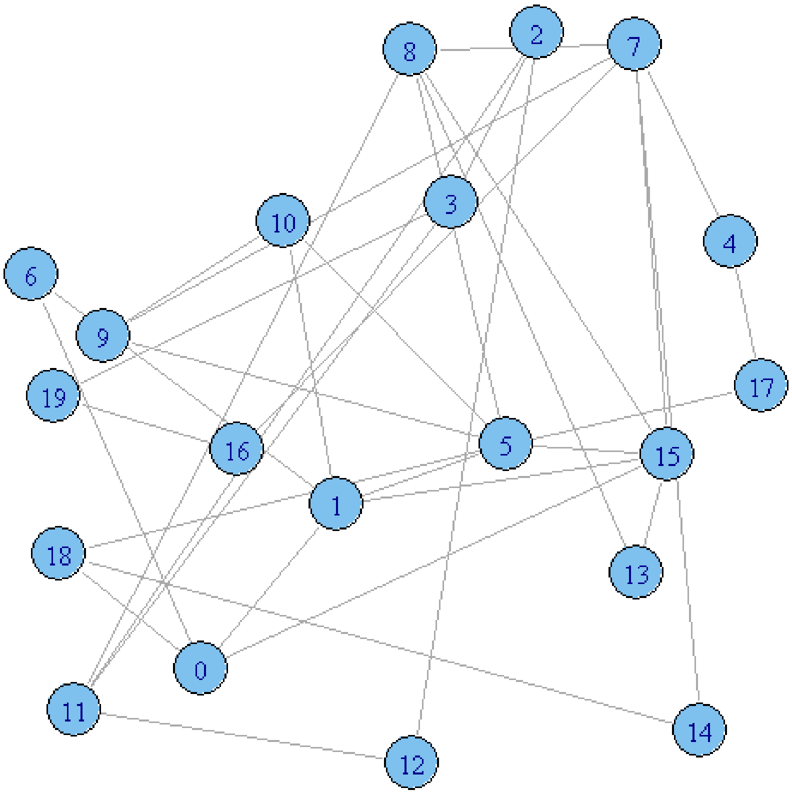


We note that, for these values of L, both α's are overestimating the active set, however α = 1/3 is still a more conservative estimate.

We have focused on lower values for L because we are more concerned in analyzing gene sets in which just a few genes are actually active, as opposed to nearly the entire set. At low values of L, α = 1/12 reported better F scores at low signals. However, it is also reporting a very large estimated active set, especially when we take correlation into account. Considering that correlation likely exists in an actual data set, we decided that to use α = 1/3.

To evaluate the performance of the adjusted P-value method in more complicated correlation structures, we designed a simulation with different levels of correlation among the active and inactive sets. We considered high correlation in the active set, with lower correlation in the inactive set. To do this, we randomly generated a network of gene interactions, given in Figure 7. We set the partial correlation of connected genes in the network in the set of first “L” genes to 0.6, and set the partial correlation of the remaining connected genes to the lower level of 0.2. In other words, instead of setting the entire active and inactive sets to be correlated, we only allow genes connected by an “edge” to be related.

Figure 7: Correlation Adjacency Matrix

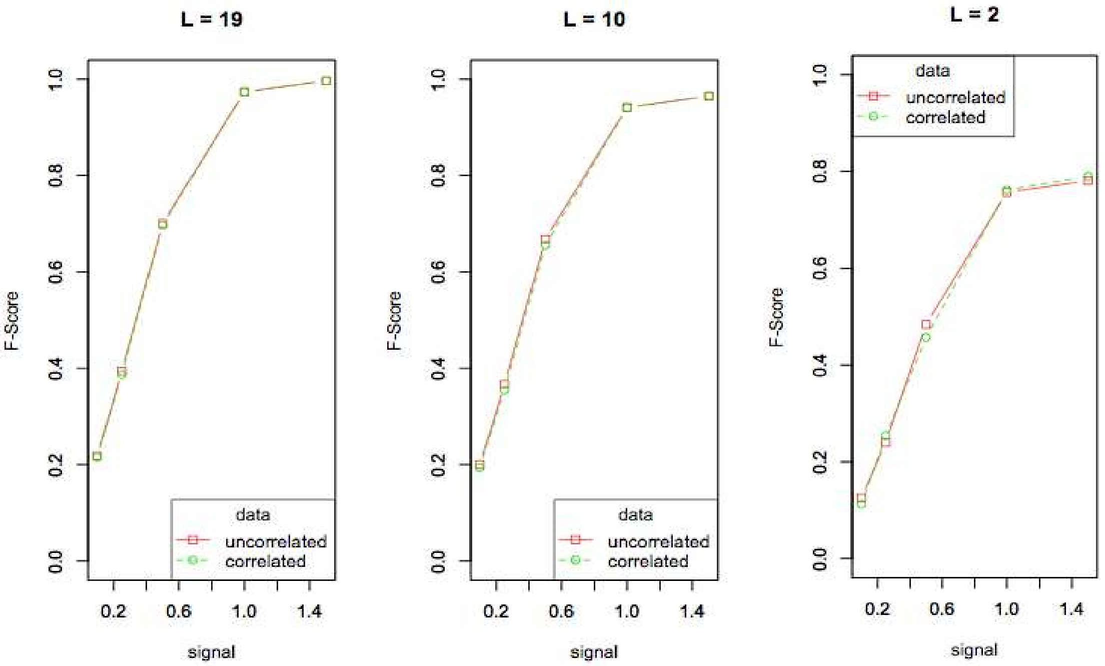


Note: Zero corresponds to the first expression level, One to the second, and so on

If two genes are connected in the graph, their expression levels would be correlated given the expression levels of the remaining genes in the graph. For instance, since zero and one are connected in the picture above, our first and second expression levels will be correlated in the sample data (for both the treatment and control). Since zero is not connected to two, their expression levels are only correlated through other genes in the graph. Using this correlation structure we, again, generated 40 sample expression levels for each of 20 genes from an i.i.d. Normal distribution with mean zero and variance one. We used the same setup for treating genes to create an active set as described before.

The simple correlation structure described before very noticeable changed the performance of our method. In certain scenarios the method performs better with correlated data than uncorrelated, and in certain scenarios it performs much worse. But this is a much more complicated structure, which we are more likely to see in actual test data. We would like our method to be robust to correlation structures, such as this, to confirm that our method will be effective in practice. Figure 8 shows the performance of our adjusted P-value method comparing F scores for uncorrelated and data generated using this new correlation structure.

Figure 8: Partial Correlation vs. Uncorrelated



In all scenarios the uncorrelated and correlated data provide almost identical results. It is quite obvious that this new correlation structure is having little to no effect on the performance of our F scores, and the adjusted P-value method is robust to these sorts of correlation structures.

3.4 Comparing With FDR

The next step is to compare the adjusted p-value method ($\alpha = 1/3$) to the Benjamini-Hochberg FDR algorithm previously described. First, a threshold for the FDR procedure should be determined. FDR, as mentioned before, works to ensure a low false positive rate. Since precision measures the proportion of the discoveries that are actually correct, we would expect the FDR procedure to maintain high precision. Setting too strict of a threshold will increase the precision, but at the expense of the recall. Figure 2 (page 10) shows F scores for thresholds of .1 and .2 using uncorrelated data. In some scenarios a threshold of .2 seems to give slightly better results, but with a low L both thresholds perform fairly equally.

Figures 3 and 4 (pages 11-13) show F scores for the FDR methods using correlated data. One interesting observation is that the simple correlation structure has the same effect on this method as on the adjusted P-values. With high correlation, the F scores for high signals drop quickly, especially with small L, while F scores for lower signals actually get slightly better. Figures 3 and 4 do not seem to provide conclusive data that either of the two thresholds is better than the other. Overall, F scores for a threshold of .2 appear to be slightly better, but we can compare magnitude of the estimated active set to show that it may still be beneficial to use a threshold of .1.

As stated before, we are more concerned with smaller values of L, as well as correlated data. In Figure 4, where L = 2 with high correlation, the F scores for both FDR thresholds appear almost identical. Figure 6 (page 16) shows the estimated active set for low L and high correlation. As discovered with our method, FDR is over-estimating the active set for small values of L. The lower threshold, .1, provides slightly more conservative estimates of the active set while retaining a comparable F score. For this reason, we will compare our adjusted p-value method to the FDR method at a threshold of .1.

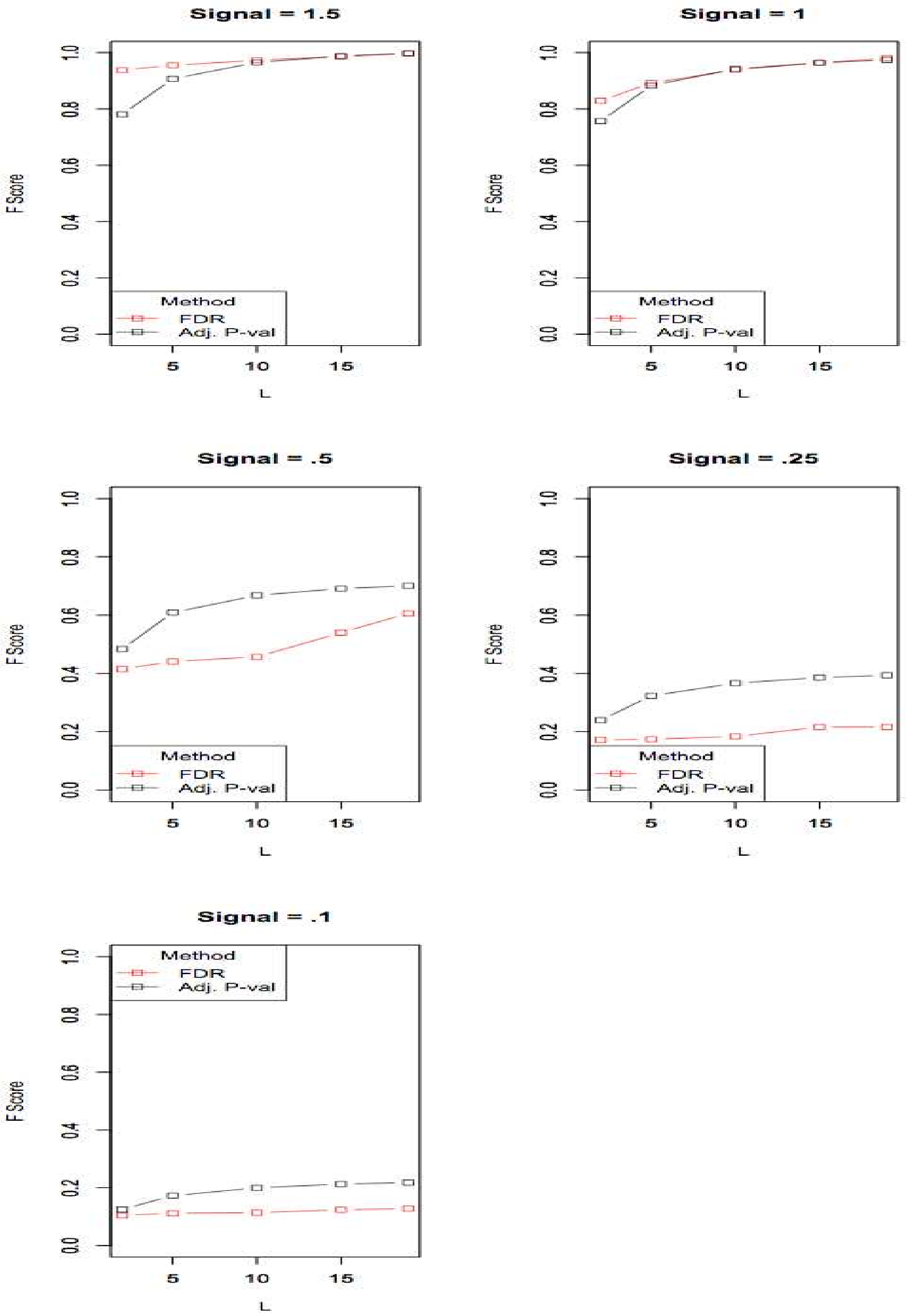
By the nature of our set up, there are a few scenarios where we can expect the methods to differ in effectiveness. When L is large there is a very small chance to identify false positives. Thus, most genes identified as active will automatically be in the active set, meaning that precision will always be high. The F score will then differ mainly due to the recall. At high signals, this should not matter because, as we have seen, both methods do very well at high values of L and high signals. At low signals, however, the FDR may be too conservative in identifying anything, and this would lower the recall. We expect our method to do better for scenarios with large L and small signal.

When L is small, the chance for identifying false positives arises simply because the set contains more inactive genes. At high signals we would still expect both of our methods to do well in identifying the truly active set, giving both methods a high recall. Since FDR focuses on avoiding false positives, it will still have high precision. Our method may not have as high a precision, meaning the FDR method will do better in these scenarios.

The interesting scenarios are those that have medium to small active sets (2, 5, 10), as well as mid to low signals (.5, .25, .1). There is the potential for false positives, as well as the potential that the FDR will be too conservative to find active genes. By fixing the signal, and altering L we can see the results best. Figure 9 shows comparisons across all 25 scenarios using uncorrelated data with $\alpha = 1/3$ and the FDR threshold set to .1 to test these predictions.

Figure 9: Adjusted P-values and FDR

Uncorrelated Data

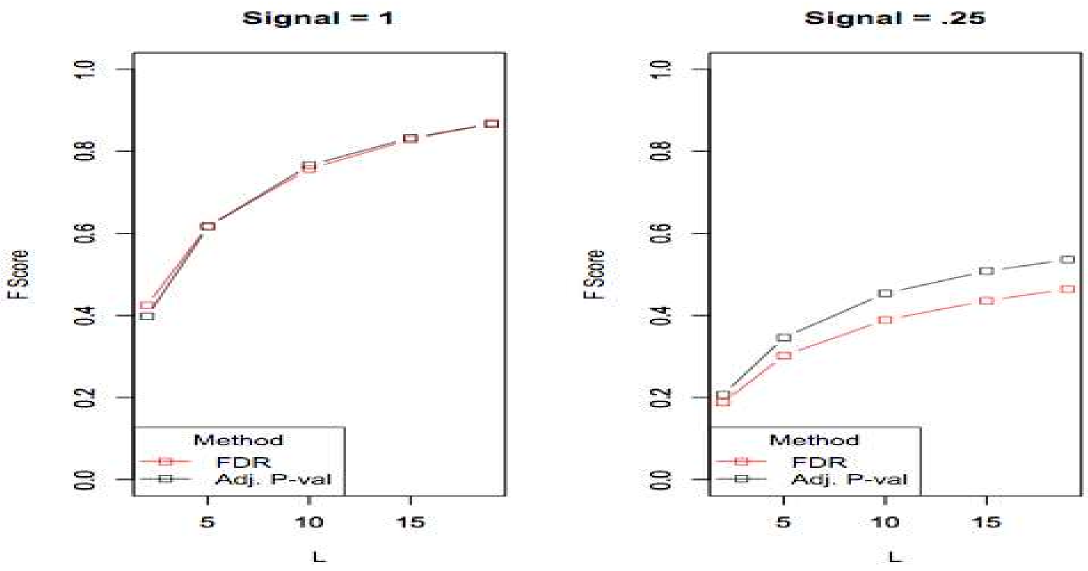


Our previous assumptions appear correct. FDR works best for large signals, but our method seems to be doing better for lower signals. More importantly, in the scenarios that we were unsure of, the adjusted P-values consistently do better than FDR. In fact, when the signal is dropped below 1, FDR never has higher F scores than our P-values. Figure 9 makes it clear that there are very few scenarios in which FDR does better than our method on uncorrelated data.

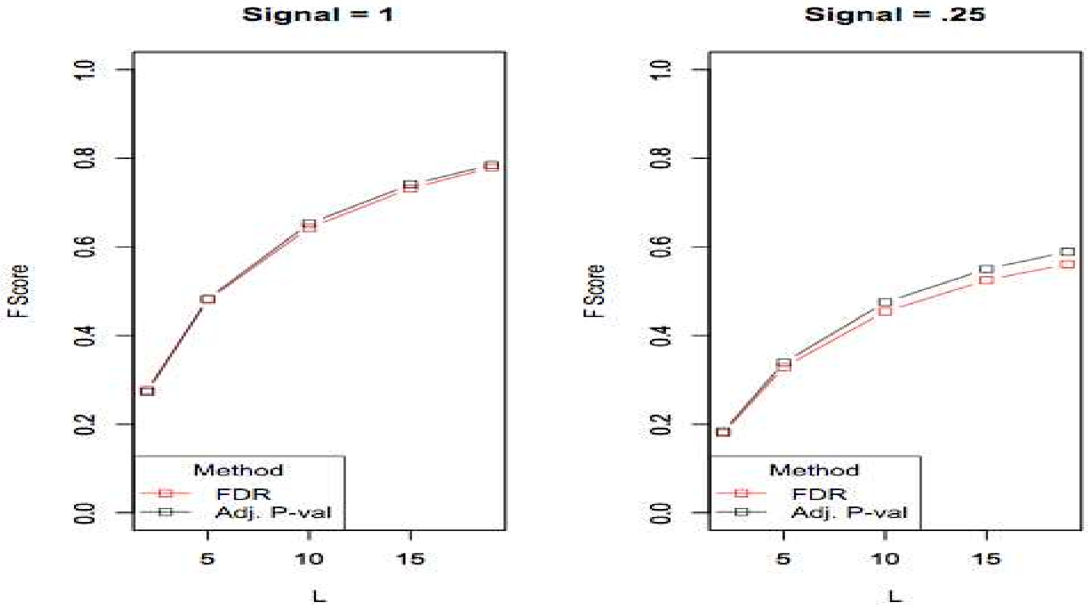
From what we have seen in previous analysis, the correlation structure appears to affect both methods in a similar fashion. As more correlation is introduced, we expect F scores for higher signals to drop, and for lower signals to rise. Figure 10 shows our results for correlations of both .2 and .8. As more correlation is introduced the F scores are consistently dropping at higher signals. At lower signals, the added correlation helps the adjusted P-value F scores slightly, but helps FDR much more. As correlation is added, the methods begin to become equally effective with regards to F values.

Figure 10: Adjusted P-values and FDR

Correlation = .2



Correlation = .8



As seen in Figures 5 and 6 (pages 15-16), FDR is consistently more conservative when estimating the active set than the adjusted P-values. We would expect the FDR method to be more conservative, but it is possible that this method of FDR is too conservative, and may be ineffective in realistic settings.

4 Analysis of Yeast Gene Expression Data

Ideker et al (2001)[5] integrated gene expression and protein level data to study significant signaling and metabolic pathways in yeast Saccharomyces cerevisiae. They reported interactions among genes and proteins in different pathways along with information on the estimated correlation among genes in the network. The authors also grouped the genes into subnetworks (pathways) based on their biological functions.

In this study, the mRNA expression levels of yeast genes are measured in 9 different strains of yeast, each including a different perturbation in one of the genes in the Galactose Utilization pathway. For each perturbation, two samples of data are available. The first set of samples represents the expression levels of genes in cells grown in presence of galactose (gal+), while the second set includes expression levels for cells grown in absence of galactose (gal--), where the main source of carbon is raffinose. Here we consider the genes in two of the pathways recognized by Ideker et al (2001), namely rProtein Synthesis and RNA processing.

We analyzed the expression levels using the FDR approach as well as our adjusted P-value method. The rProtein Synthesis and RNA processing pathways consist of 28 and 78 genes, respectively. The FDR approach was very conservative, only discovering one significant gene in each pathway. Our method, however, discovered 13 significant genes in each pathway. Table 2 summarizes the result from the adjusted p-value method, listing the genes with the smallest P-values first.

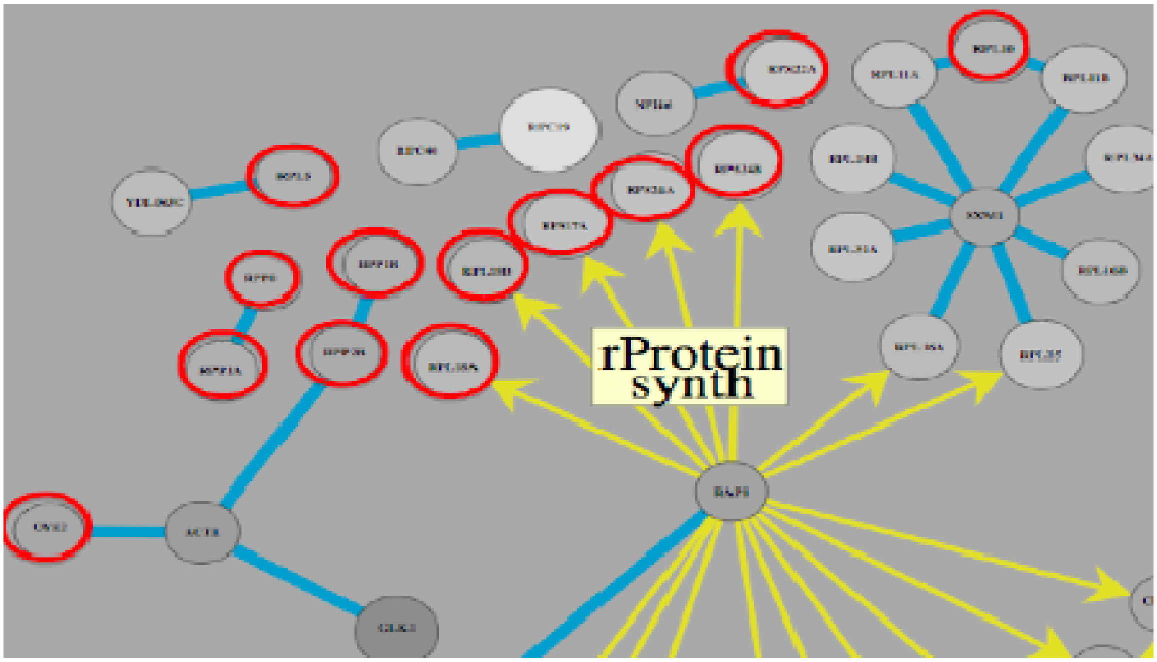
Table 2: Test Data Results

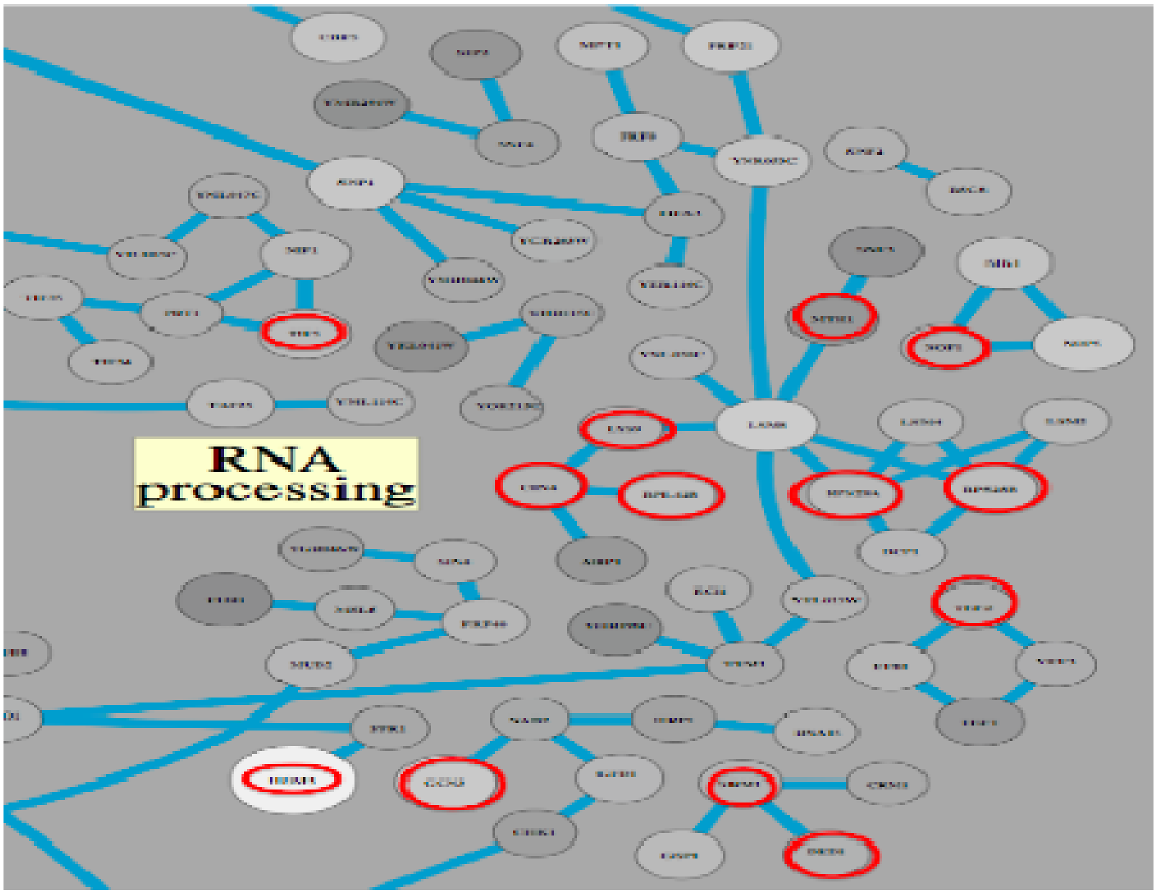
Rank	**rProtein Synthesis**	**RNA Processing**
1	RPL10*	RPL42B*
2	RPP1A	SRM1
3	RPP2B	NOP1
4	RPS22A	RPS28B
5	RPP1B	LYS9
6	OYE2	DED1
7	RPL18B	CIN4
8	RPP0	RPS28A
9	RPS17A	GCN3
10	RPS24B	HOM3
11	RPL5	TEF2
12	RPS24A	TIF5
13	RPL18A	MTH1

* Denotes that FDR also found this gene to be significant.

Interpreting these results can be a bit difficult, as we are unsure of how many genes are truly active. In our sample data we determined the active set, so we knew which genes we should be declaring active, but we do not know this here. The FDR approach is conservative, but it could be true that there are very few, if any, active genes in each pathway. We can gain a little more information by observing the spatial arrangement of each pathway. We would hope that the genes we found to be active will be near each other in this arrangement, indicating a relationship. Figure 10 shows the arrangement of each pathway, with the genes we found to be active highlighted in red.

Figure 10: The rProtein Synthesis and RNA Processing Pathways





In the RNA Processing pathway there appears to be two clusters of genes, one in the middle and one at the bottom, that we considered active. There are a few gaps between some of the genes within the clusters, though, which may weaken the argument that those genes are the true active set. We see a much more interesting pattern within the rProtein Synthesis Pathway. We see that we declared RPL18A, RPL18B, RPS17A, RPS24A, and RPS24B all to be active, and that they are all controlled by the same transcription factor. Further, most of the other genes declared active are clustered near these five genes. While we cannot say for sure that these genes are truly active, we know that they are closely related, and if one of them is active it is more likely that the rest are active as well.

5 Conclusions and Future Work

With our simulated data, the adjusted P-value method always works comparable to, and in many cases preferable to, the FDR approach. With the test data, we found the FDR procedure to be very conservative in reporting genes as active, while the adjusted P-value method shows the opposite behavior. While some of our results from the test data seem promising for our method, it is difficult to claim that the adjusted P-values perform better than FDR. It would be interesting to investigate such a huge discrepancy more systematically.

Some approaches to FDR account for correlation structures using data permutations. To do these permutations effectively, however, there must be a large number of genes in a pathway. Since our simulated data only focused on pathways of twenty genes, we were not able to properly compare our method permutation-based FDR methods. It is possible that more complicated methods of FDR will perform better than the Benjamini-Hochberg (1995) method that we compared our method to. It would be beneficial to simulate data with much larger pathways in order to make this comparison.

We do believe that we might be able to improve the adjusted P-value method. When deciding which alpha (1/12, 1/6, 1/4, 1/3) works best, it was clear that which worked best depended on which scenario we looked at. Sometimes alpha = 1/12 worked better, and sometimes alpha 1/3 worked better. It may be beneficial to adaptively estimate alpha according to the data we are looking at. Namely, we would like to perform some re-sampling technique in order to determine alpha based on the correlation structure of the data. This should improve our method, and might be necessary to make it competitive with permutation-based methods of FDR.

6 Acknowledgments

I would like to thank Drs. Moulinath Banerjee, George Michailidis, and Ali Shojaie of the Department of Statistics at University of Michigan for their invaluable guidance and help in designing the methods, and furthering my understanding of the complex concepts involved.

References

[1] Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. Annals of Applied Statistics, 1(1):107–129.

[2] Subramanian, A. and Tamayo, P. Mootha, V. K. and Mukherjee, S. and Ebert, B. L. and Gillette, M. A. and Paulovich, A. and Pomeroy, S. L. and Golub, T. R. and Lander, E. S. and Mesirov, J. P. (2005). A knowledge-based approach for interpreting genome-wide expression profiles. PNAS. 102, pg 15545-15550.

[3] Storey, J. (2010) False discovery rates. In International Encyclopedia of Statistical Science, Lovric M (editor)

[4] Benjamini, Y., Hochberg, Y. (1995) “Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing” Journal of the Royal Statistical Society, Series B (Methodological) **57** (1): 125–133.

[5] Ideker, T., Thorsson, V., Ranish, J., Christmas, R., Buhler, J., Eng, J., Bumgarner, R., Goodlett, D., Aebersold, R., and Hood, L. (2001). Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. Science, 292(5518):929.