ANALYZING MICROARRAY TIME SERIES DATA USING A HAAR WAVELET TRANSFORMATION

by

Sarah D. Braudrick

B.A. Mathematics, B.A. Sociology, University of Northern Colorado, 1998

A thesis submitted to the University of Colorado at Denver and Health Sciences Center in partial fulfillment of the requirements for the degree of Master of Science Mathematical Sciences 2007
This thesis for the Master of Science
degree by
Sarah D. Braudrick
has been approved
by

[Signature]
Stephen C. Billups

[Signature]
Kathe E. Bjork

[Signature]
Stephan R. Sain

4/18/07
Date
Braudrick, Sarah D. (M.S., Applied Mathematics)
Analyzing Microarray Time Series Data Using a Haar Wavelet Transformation
Thesis directed by Associate Professor and Graduate Program Director Stephen C. Billups

ABSTRACT

This thesis presents an algorithm for identifying significant temporal patterns in microarray time series. Using a false discovery rate procedure to assess the magnitude of wavelet coefficients resulting from a Haar wavelet transformation, the algorithm detects gene expression signals exhibiting time-dependent features. Numerous simulated time series, each consisting of a small proportion of temporally dependent functions among random signals, were generated by which to test the algorithm. The simulation encompassed the following changeable scenarios: number and distribution of time points (evenly-spaced or randomly-spaced), noise-to-signal ratio, and level of frequency. Testing the algorithm using these parameters, we were able to answer questions regarding the effect of sample quantity, sampling rate, and sample quality (degree of noise present in the expression signal readings) on the algorithm's performance in detecting temporal patterns in a microarray time series. Finally, the algorithm is applied to a genome-wide microarray time series collected from Rattus norvegicus mammary tissue over the course of the 4-day estrous cycle.
This abstract accurately represents the content of the candidate’s thesis. I recommend its publication.

Signed

Stephen C. Billups
DEDICATION

This thesis is dedicated to John, for the support and the space and knowing when to give which, and to my parents, who never ran out of encouragement.
ACKNOWLEDGMENT

I'd like to thank Steve Billups for giving me a thesis topic which was as interesting as it was challenging, and for always giving thorough feedback. And also thanks to Holly Wolf, for her limitless understanding when it came to work and school conflicts.
CONTENTS

Figures ................................................................................. x
Tables ..................................................................................... xvi

Chapter
1. INTRODUCTION ................................................................. 1
1.1 Time Series Gene Expression Data Analysis ...................... 1
1.2 Related Work ................................................................. 1
1.2.1 Early attempts - time unordered ................................. 2
1.2.2 Recent attempts - time ordered ................................. 2
1.2.3 Wavelets in signal processing ................................. 3
1.3 Our Approach ................................................................. 3
1.3.1 Motivation ................................................................. 4
1.4 Challenges ................................................................. 5
1.4.1 Sampling ................................................................. 6
1.4.2 Homogeneity .......................................................... 6
1.4.2.1 Within samples ................................................. 7
1.4.2.2 Between samples ............................................. 7
1.4.3 Replicates ............................................................... 8
1.5 Conceptual Framework ............................................... 8
1.5.1 A model of the data .............................................. 9
1.6 Implementation .......................................................... 10
4.2 Variation of Sample Response ...................................... 70
4.3 Sampling Rate............................................................ 70
4.4 Frequency of the Underlying Temporal Response .............. 71
4.5 Which Wavelet Coefficient Performs Best? ...................... 72
5. FUTURE WORK .................................................................. 73
References ........................................................................... 75
LIST OF FIGURES

Figure

2.1 The eighth-order Haar wavelet basis functions on the interval $[0, 4]$. Here, $n = 8$ because $D = 2$. .................... 16

3.1 The Haar Wavelet step function, $\psi$ ................................. 28

3.2 Sinusoidal functions of various frequencies on the time interval $[0,1]$. ..................... 29

3.3 Regression functions 1 through 4, on the interval $[0,1]$. ...................... 30

3.4 Surface plots of observed sensitivity rates. (These are 3-dimensional renderings of Tables 3.2 and 3.3, viewed from their upper left hand corners towards their lower right.) ................................. 32

3.5 Sensitivity rates for $16 \geq T \geq 9$ with $\eta = 0.1$. ..................... 33

3.6 Haar wavelet-generated (normalized) signals for $T = 40$, where $T$ is evenly spaced and $p = 0.1$. This illustrates how, even with a large number of uniform samples, true signals (black) can quickly become nearly indistinguishable from pure noise (grey) when $\eta$ increases from 1.5 to 2.0. ......................... 34

3.7 Surface plots for the observed FDRs from uniform and randomly spaced $T$, $p = 0.1$. ................................. 35

3.8 Surface plots of observed sensitivity when the proportion of true signals ($p$) is 50% (a and b) and 90% (c and d). No significant differences were seen between small and large $p$. ...................... 36

x
3.9 Mean sensitivity rate across all evenly spaced $T$ (with $\eta = 0.1$) as a function of frequency rate. .................................................. 37
3.10 Surface plots of the resulting sensitivity from low frequency (near 0 and 1) sinusoidal-generated simulations. ........................................... 38
3.11 Surface plots of the resulting sensitivity from lower frequency (near 2 to 4) sinusoidal-generated simulations. ................................. 39
3.12 Surface plots of the resulting sensitivity from higher frequency (greater than 5) sinusoidal-generated simulations. .............................. 40
3.13 Sensitivity rates for $\text{frq} = 0.3, \eta = 0.1$, as a function of $T$ in finer detail. Sensitivity decreases more sharply for evenly spaced $T$ (solid dark line) between 20 and 15 than for randomly spaced $T$ (dashed light line). .................................................. 42
3.14 Sensitivity rates for $\text{frq} = 0.3, T = 50$, as a function of $\eta$ in finer detail. Sensitivity decreases most when $0.5 \leq \eta \leq 1.0$. Evenly (solid dark line) and randomly (dashed light line) spaced large $T$ produce similarly decreasing rates. .................................................. 43
3.15 Sensitivity rates for $\text{frq} = 1.3, \eta = 0.1$, as a function of $T$ in finer detail. Sensitivity decreases more sharply for evenly spaced $T$ (solid dark line) between 30 and 25 than for randomly spaced $T$ (dashed light line). .................................................. 45
3.16 Sensitivity rates for $\text{frq} = 1.3, T = 50$, as a function of $\eta$ in finer detail. Evenly (solid dark line) and randomly (dashed light line) spaced large $T$ produce similarly decreasing rates. .................................................. 45
3.17 Sensitivity rates for freq = 1.8, \( \eta = 0.1 \), as a function of \( T \) in finer detail. Sensitivity decreases more sharply for evenly spaced \( T \) (solid dark line) between 14 and 12 (see arrows) than for randomly spaced \( T \) (dashed light line). 

3.18 Sensitivity rates for freq = 1.8, \( T = 50 \), as a function of \( \eta \) in finer detail. Evenly (solid dark line) and randomly (dashed light line) spaced large \( T \) produce similarly decreasing rates and both have a high tolerance to a large amount of added noise.

3.19 Sensitivity rates for freq = 2.5, \( \eta = 0.1 \), as a function of \( T \) in finer detail. Sensitivity for evenly spaced \( T \) (solid dark line) is larger when \( T = 24 \) than when \( T = 29 \), and a large decrease occurs between \( T = 24 \) and 23. For randomly spaced \( T \) (dashed light line), sensitivity is larger when \( T = 22 \) than when \( T = 31 \).

3.20 Sensitivity rates for freq = 2.5, \( T = 50 \), as a function of \( \eta \) in finer detail. Evenly (solid dark line) and randomly (dashed light line) spaced large \( T \) produce similarly decreasing rates except for \( 0.5 \leq \eta \leq 0.7 \), where randomly spaced \( T \) yields smaller sensitivity rates.

3.21 Sensitivity rates for freq = 3.8, \( \eta = 0.1 \), as a function of \( T \) in finer detail. Sensitivity drops significantly, but uniformly, for evenly spaced \( T \) (solid dark line) between 24 and 20 (see arrows). Randomly spaced \( T \) (dashed light line) yields slightly higher rates for smaller values of \( T \).
3.22 Sensitivity rates for \( \text{frq} = 3.8, \, T = 50 \), as a function of \( \eta \) in finer detail. Evenly (solid dark line) and randomly (dashed light line) spaced large \( T \) produce similarly decreasing rates.

3.23 Surface plots (b and c) of observed sensitivity rates from simulations generated from regression function (a) \( f_{\text{reg}r1}(t) = 1 - 48t + 218t^2 - 315t^3 + 145t^4 \).

3.24 Sensitivity rates for \( f_{\text{reg}r1}(t), \, \eta = 0.1 \), as a function of \( T \) in finer detail. Evenly spaced \( T \) (solid dark line) differs most from randomly spaced \( T \) (dashed light line) near \( T = 30 \). Rates decrease significantly for both once \( T \) reaches the mid-twenties.

3.25 Sensitivity rates for \( f_{\text{reg}r1}(t), \, T = 50 \), as a function of \( \eta \) in finer detail. Evenly (solid dark line) and randomly (dashed light line) spaced large \( T \) produce similarly decreasing rates except for \( 0.4 \leq \eta \leq 0.6 \), where randomly spaced \( T \) yields smaller sensitivity rates.

3.26 Surface plots (b and c) of observed sensitivity rates from simulations generated from regression function (a) \( f_{\text{reg}r2}(t) = 0.3e^{-64(t-0.25)^2} + 0.7e^{-256(t-0.75)^2} \).

3.27 Surface plots (b and c) of observed sensitivity rates from simulations generated from regression function (a) \( f_{\text{reg}r3}(t) = 10e^{-10t} \).

3.28 Surface plots (b and c) of observed sensitivity rates from simulations generated from regression function (a) \( f_{\text{reg}r4}(t) = e^{k(t-\frac{1}{3})} \), where \( k = -2 \) for \( t < \frac{1}{3} \) and \( k = 1 \) for \( t \geq \frac{1}{3} \).
3.29 Sensitivity rates for \( f_{\text{reg}4}(t) \), \( \eta = 0.1 \), as a function of \( T \) in finer detail. Rates for randomly spaced \( T \) (dashed light line) decrease more gradually than evenly spaced \( T \) when the number of time points is less than 15. And is evenly spaced \( T \) favoring an even number of time points when \( T \) is small? .................................................. 59

3.30 Sensitivity rates for \( f_{\text{reg}4}(t) \), \( T = 50 \), as a function of \( \eta \) in finer detail. Evenly (solid dark line) and randomly (dashed light line) spaced large \( T \) produce similarly decreasing rates except for \( 0.7 \leq \eta \leq 1.0 \), where randomly spaced \( T \) yields smaller sensitivity rates. .................. 59

3.31 A plot indicating the time points sampled. Each asterisk marks the time (in hours) that a tissue sample was taken from a rat specimen across the 4 day estrous cycle. .................. 60

3.32 The Haar wavelet reconstruction of the normalized expression signal for APeg1. An increase in expression appears in the estrus and metestrus phases. .................. 62

3.33 The Haar wavelet reconstruction of the normalized expression signal for RMT-1. Expression values are high in the first half of the estrous cycle, and low in the second half. .................. 63

3.34 The Haar wavelet reconstruction of the normalized expression signal for IL-17B. Expression is low in the first half of the cycle, and high during the last two phases. .................. 64

3.35 The Haar wavelet reconstruction of the expression signal for TSP2. Expression is higher in the first half, and lower in the second half of the estrous cycle. .................. 65
3.36 The Haar wavelet reconstruction of the expression signal for LAMA5.

Expression moderately decreases from the beginning to the end of the cycle, with some high level readings occurring in the estrus phase.
LIST OF TABLES

Table

2.1 FDR procedure pseudocode ........................................ 22
2.2 Algorithm pseudocode ............................................. 23
2.3 Algorithm input parameters ....................................... 24
3.1 Input parameters needed to construct $F$ ......................... 27
3.2 Observed sensitivity rates for $f = \psi(t)$, evenly spaced $T$ and $p = 0.1$ 31
3.3 Observed sensitivity rates for $f = \psi(t)$, randomly spaced $T$ and $p = 0.1$ 32
3.4 Observed sensitivity rates for $f = \sin(0.3\pi t)$, $T$ evenly spaced and $p = 0.1$ 41
3.5 Observed sensitivity rates for $f = \sin(1.3\pi t)$, evenly spaced $T$ and $p = 0.1$ 44
3.6 Observed sensitivity rates for $f = \sin(1.8\pi t)$, evenly spaced $T$ and $p = 0.1$ 46
3.7 Observed sensitivity rates for $f = \sin(2.5\pi t)$, evenly spaced $T$ and $p = 0.1$ 48
3.8 Observed sensitivity rates for $f = \sin(3.8\pi t)$, evenly spaced $T$ and $p = 0.1$ 50
3.9 Observed sensitivity rates for $f_{\text{reg}1}(t)$, evenly spaced $T$ and $p = 0.1$ 53
3.10 Observed sensitivity rates for $f_{\text{reg}2}(t)$, randomly spaced $T$ and $p = 0.1$ 55
3.11 Observed sensitivity rates for $f_{\text{reg}3}(t)$, evenly spaced $T$ and $p = 0.1$ 56
3.12 Observed sensitivity rates for $f_{\text{reg}4}(t)$, evenly spaced $T$ and $p = 0.1$ 57
3.13 A frequency distribution of significance rates for 10 applications of the algorithm using $R = 5$ ............................................. 61
3.14 Genes found to have temporal significance 100% of the time ... 67
4.1 Percentage Distribution of Positive-Responsible $C_k$ (when $T = 50$ and $\eta = 0.1$) ................................................................. 72
1. INTRODUCTION

1.1 Time Series Gene Expression Data Analysis

Microarray technology allows researchers to quantify the expression levels of thousands of genes within a tissue sample. By measuring mRNA levels associated with each gene, microarrays provide a snapshot of the genetic activity present in that tissue. This information can then be used to understand patterns of gene expression. Researchers have used microarrays to investigate relationships among genes, how they react to environmental changes – such as injury, disease, and drugs – and their roles in the routine maintenance of an organism. Gene function plays a dynamic role in the regulation of biological processes, and time series analysis can help us understand how this unfolds. Microarray experiments, when performed successively, can reveal temporal patterns in biological systems such as cellular cycles, organism development, and disease progression. The resulting data from these time course experiments, however, can be difficult to manage; they tend to be enormous in number (often comprising an entire genome), widely variable, and diverse in scope (some profiles averaging mRNA levels near 10,000, others near 500). The analysis of this complex data introduces unique challenges to investigators and calls for specialized techniques that take into account the sequential nature of the samples.

1.2 Related Work

1.2.1 Early attempts - time unordered
Initial efforts to make sense of microarray time series data largely ignored the continuous nature of time, instead treating time as a categorical variable. Many early undertakings sought only to differentiate among groups of genes in the hopes of gaining a more sophisticated understanding of biological processes. Groupings tend to be established either through classification schemes, in which genes are categorized based on some pre-specified criteria, or through clustering methods which identify naturally occurring similarities. Early approaches in classifying genes were largely guided by multivariate statistical and computational methods including classical statistical modeling techniques [34, 37], clustering algorithms [12, 30], principal component analysis (PCA) with singular value decomposition (SVD) [3, 14], self-organizing maps (SOM) [28], and support vector machines [8]. The shortcoming of a multivariate analysis framework, however, is that the temporal ordering of samples is inconsequential; any permutation of time points will not change the results of the analysis.

1.2.2 Recent attempts - time ordered

More recent classification techniques have addressed this problem by recognizing that temporal order and the intrinsic continuity of gene expression signals must be incorporated. Some of these time sensitive classification schemes include the use of self-organizing maps adjusted to incorporate a temporal metric [27, 36], time warping algorithms that attempt to co-align expression signals [1], and functional data analysis (FDA) methodologies that view observed gene profiles as independent components in a smooth stochastic process [22, 35]. The latter presented methods specifically for unevenly and/or infrequently sampled data. Notable nonparametric methods presented for modeling time series ex-
pression profiles have included the use of splines [5, 11] and wavelets. In work
done by Klevecz in 2000 and Klevecz and Murray in 2001 [17, 18], wavelets were
used to identify dominant frequencies at which large numbers of genes exhib-
ited oscillations in their expression profiles. Our use of wavelets is qualitatively
different. Specifically, we will use wavelets as a means of identifying individual
genomes that exhibit significant time-dependent variation in their expression levels.

1.2.3 Wavelets in signal processing

The use of wavelets in time series analysis is not new. Wavelets have been
applied to a variety of series including the detection of time-dependent climate
signals [20], serial dependencies within functional magnetic resonance imaging
(fMRI) [9], the analysis of electrocardiogram (EKG) heartbeat intervals [16] and
electroencephalogram (EEG) epileptic spikes [19]. In image processing, wavelets
have been used in the analysis of 2-dimensional images (which are just signals
arising from spacial dependencies instead of temporal), to identify patterns at
various resolutions. Examples include JPEG image compression [31], optimiza-
tion of image coding [33], edge detection and the removal of noise in images [24].

1.3 Our Approach

Here we present a method for identifying meaningful temporal patterns in
gene expression microarray data. Our algorithm computes the Haar wavelet
transformation on the expression profile of each gene, and then incorporates
a false discovery rate analysis on the decomposition coefficients to single out
genomes with significantly large temporal responses, relative to the noise. Prior to
describing the details of the algorithm, it will be useful to first describe the data set that motivated the development of this method.

1.3.1 Motivation

The development of our algorithm was motivated by the desire to analyze a genome-wide time series that resulted from research investigating the estrous cycle. The data came from microarray experiments measuring the levels of mRNA present in mammary tissue samples taken from female rats (Rattus norvegicus) and spanning the course of the rat estrous cycle (typically 4-5 days). The experiment, conducted by Margaret C. Neville and Pepper Schedin at the University of Colorado's School of Medicine, sampled 32 time points over a four day period at approximately 90 minute intervals between the hours of 7am and 7pm. The microarrays were set up to detect expression levels for the entire rat genome, comprising over 21,000 genes.

Much of the genetic architecture underlying the estrous cycle is not clearly understood, therefore our quest was exploratory in nature. The hope was to be able to identify a set of candidate genes exhibiting a significant temporal response worthy of further investigation in future experiments. Features of this experiment made the prospect of data analysis both encouraging and daunting. While there were many time points sampled, and with some level of regularity in sampling rate, there were also large gaps of time with no samples, coupled with the additional variability introduced by heterogenous sample subjects.

Unlike some experiments in which the samples are homogenous, (e.g., the time points are sampled from cloned organisms or from the same individual multiple times), this experiment required the extraction of vital organs and
therefore each time point represented a different rat. The lack of sample homogeneity presents uncertainty that can complicate analysis efforts because each rat's genetic makeup varies to some unknown degree. In addition to this variability between individual rats, there can also be disparities within samples. For example, the presence of peripheral tissue whose primary biological function may be far different than that of the tissue under examination can result in expression values reflecting a different process. The heterogeneity of samples also raises concern for how much confidence we can have that the rat sacrificed at time $t_k$ was a good representation of the estrous cycle at time $t_k$. In other words, how can we be sure that the rat representing time point $t_7$, let's say, wasn't actually better-suited to represent time point $t_9$ in the estrous cycle?

The sampling rate of this set had irregularities resulting in large windows of unmeasured gene activity (over nights), therefore we wanted to explore how uneven sampling might affect the identification of significant temporal profiles. An integral part of our testing scheme involves the comparison between uniformly and randomly spaced time points.

Encouraging however, was that this set had a relatively large number of samples—a luxury given that, due to the high cost involved, it is not uncommon for microarray time courses to sample less than 10 time points [13]. While the experiment did not perform replicates, this forfeiture afforded the advantage of more time points covered. Also attractive is the regularity of sampling that did occur within the four 12 hour sampling phases.

1.4 Challenges
The idea behind gene expression time series experiments is to obtain enough snapshots of gene activity to reveal temporal patterns within a particular biological process or stimulus response. A key factor in accomplishing this is being able to distinguish between the true signal and its accompanying noise. As will be described, there are many factors in the experimental design that can complicate the discernment between pure signal and noise.

1.4.1 Sampling

As with many time course experiments, sampling methodology can greatly affect the subsequent assessment of data. The rate of sampling should be tailored to the process being studied; longer intervals may accommodate slowly unfolding gene activity, whereas shorter intervals become necessary when studying intense biological responses. While sampling frequently can simplify and strengthen analysis, it may be infeasible due to the high cost of microarrays. Sampling too infrequently, however, can make it nearly impossible to interpolate gene activity between sample points.

Even when many time points are sampled, irregularity of sampling intervals can complicate analysis because the rate of change between time points increases as the interval decreases. Sometimes irregular intervals are strategically carried out in order to capture particular phases of a process where rapid or slow behavior is expected [26]. But in many situations, an uneven rate of sampling is difficult to avoid due to the impracticality of around-the-clock sampling.

1.4.2 Homogeneity

In addition to sparse or irregular sampling, other factors may bring uncertainty into the analysis, namely, inconsistencies within and between tissue
samples.

1.4.2.1 Within samples

Inconsistencies within samples occurs when the extracted specimen contains tissue from multiple organs. For the purposes of many experiments, the prototypical tissue sample contains only tissue from the organ of interest, be it heart, liver, or skin, etc. This is ideal because gene activity varies among tissue types depending upon their biological functions. For example, certain genes responsible for regulating functions of sight will be highly expressed in ocular tissue, yet nearly dormant in proximal brain tissue. But during extraction, samples can be inadvertently contaminated with tissue from neighboring organs.

1.4.2.2 Between samples

Inconsistencies between tissue samples is the result of using heterogenous subjects. A favorable experimental condition is when each tissue sample comes from the same organism, or a clone. Such control is attainable when working with populations of smaller organisms like C. elegans or Saccharomyces cerevisiae (yeast) because lines of these organisms can be perpetuated indefinitely. But often it is necessary to use organisms more similar to humans, in which case cloning is not (yet) a practical approach and each tissue sample requires the sacrifice of a unique organism. When each time point represents a unique individual, it can be difficult to know the degree to which changes in expression values are attributable to a biological response or to the peculiarities of the individual. Measures can be taken to align organisms so that each sacrificed specimen can reasonably represent the proper phase of the cycle for that time
point. But there will still be variation resulting from the uniqueness of the samples and which is indistinguishable from the changes inherent in the true gene signal.

1.4.3 Replicates

In traditional types of time series, the value at time point $t_j$ is directly influenced by the observation at time point $t_{(j-1)}$. This characteristic can bring robustness to the validity of the value at time point $t_j$; when the value at time point $t_j$ is similar to its neighbors, we have more confidence in it. In gene expression time series on the other hand, individual time points are independent of the preceding observations. Therefore, it is desirable to take measures to increase confidence in the value at time point $t_k$.

A common experimental design, intended to substantiate the expression value associated with each sampled time point, is to generate replicates. Replicates are additional microarray tests performed on each sample and combined into a single reading for that time point. While this design can increase confidence in the values, it does not necessarily always behoove the researcher to adopt this setup. Consider an experiment in which one is investigating a slow process over a large cycle length; requiring replicates may amount to squandering expensive microarray tests that would be more useful if spread out to cover more individual time points.

1.5 Conceptual Framework

We recognize that significant temporal patterns are not necessarily cyclical. In fact – as evidenced when one views a plot of expression values against time – what can be seen is a great deal of seemingly erratic spikes and bursts of activity.
Wavelet transformations are well-suited to such data since they are capable of closely approximating highly irregular, chaotic signals.

1.5.1 A model of the data

In designing a model of this type of data, we set forth some expectations. In general, we assume that the variation in gene expression signals is governed by three components: the temporal response, the sample response, and random noise. We assume that the temporal response component represents the time-dependent signal of the true genetic profile and that it is a continuous function of time.

The sample response is considered time-independent variation occurring from sample to sample (e.g., differences between individual organisms or tissue makeup). Note that the sample response may exhibit correlations among genes; there may be groups of genes that vary similarly from one sample to the next. The magnitude of the sample response is difficult to predict; it could be very large or small relative to the temporal response.

The random noise is also considered to be time-independent, but largely gene-independent and relatively small compared to the other two components. If we combine the sample response and the random noise into a single component, $\varepsilon$, then our model, $f$, can be written as

$$f = y + \varepsilon, \quad (1.1)$$

where $y$ represents the temporal response, i.e., the genetic profile.

Questions that we explore using our algorithm on strategically simulated data include whether the sample response is proportional to the temporal re-
sponse and what type of distribution might be representative of the sample response.

1.6 Implementation

This paper presents a method to identify significant temporal responses among a large number of gene expression signals. We model the data via a Haar wavelet transformation and use features of the reconstructed signal to identify profiles with meaningful temporal patterns. Issues to be addressed in our approach include the questions of how many samples are necessary within a given time interval, and to what degree irregularity in sampling rate can be tolerated in order to successfully detect temporal patterns. Also investigated is how best to normalize the raw data prior to the wavelet transformation. Normalization of the data is an important issue because it has the potential of magnifying small scale elements as well as minimizing those of large scale.

To test our methodological answers to these questions, we apply our algorithm to a variety of simulated datasets. After refining our algorithm during the testing phases, we apply the final version to the set of genome-wide expression levels sampled from a population of female rats in an attempt to explore the regulation of the estrous cycle.
2. METHODS

2.1 Overview of the Method

Our methods were based upon the model that gene expression measurements are made up of three components: the temporal response, a continuous function of time that reflects the true expression value; the sample response, a time-independent form of variability that results from differences between samples; and noise, which is presumed to be independent of both time and sample. Furthermore, we assume that the non-temporal portion of the signal (the sample response combined with the random noise) tends to be proportional to the signal’s temporal response.

We are primarily interested in identifying genes that have a significant temporal response. We model the temporal response for a gene as a function of time, which we approximate from the expression data. To make this concrete, consider a set of $T$ ordered data points, $(t_1, y_1), \ldots, (t_T, y_T)$, believed to have been sampled with noise from some underlying function of time. The aim is to find a new function, $f$, that closely approximates, or fits, this underlying function such that for each $j \in [1, T]$

$$f(t_j) \approx y_j. \quad (2.1)$$

Our approach models the data as a continuous function of time so that the underlying function, $y$, is approximated by the following linear model:

$$f(t) = \sum_{k=1}^{n} c_k \phi_k(t), \quad (2.2)$$
where $\phi_1(t), \ldots, \phi_n(t)$ is a family of basis functions for our method, which target temporal features we are interested in. In this study, we will use Haar wavelet functions, defined in 2.3, to form the basis. We then minimize the model’s least-squares residual to choose the best parameters $c_k$. That is, we minimize $\sum_{j=1}^{T} (y_j - f(t_j))^2$. The result of this process is a representation of each gene by the vector of coefficients $c = [c_1, \ldots, c_n]$ corresponding to the least squares fit. If the magnitude of any of these coefficients is large relative to the other coefficients, we say that the gene has a significant temporal response with respect to the corresponding basis function. A main contribution of this thesis is an algorithm for determining when a coefficient is large enough to be considered significant. The remainder of this chapter describes the algorithm in detail.

The raw dataset, $\hat{G}$, is arranged into an $m \times T$ matrix such that $\hat{G}_{i,j}$ is the expression level at the $j^{th}$ time point for the $i^{th}$ gene, where $j = 1, 2, \ldots, T$ and $i = 1, 2, \ldots, m$. Prior to calculating the functional fit, we first normalize the data, resulting in an $m \times T$ matrix $G$. The normalization process is described in Section 2.2. Section 2.3 describes the Haar wavelets, which will be used to define the basis functions $\phi_k(t)$ for equation (2.1). Section 2.3 also gives the mathematical details for calculating the least squares approximation. Section 2.4.1 describes the method for determining the significance of the magnitudes of the wavelet coefficients.

2.2 Normalization

We examined a number of methods for standardizing the data prior to performing the wavelet decomposition. The choice of a normalization scheme depends on several factors and therefore it was important to assess the implica-
tions of each standardizing component as it effects the analysis of the data. Our method was to first take the log-transform of the data and then center each signal about 0.

In microarray data analysis it is standard practice to work with the logarithm of the data instead of the raw expression values. This is primarily because of the wide scope of the expression readings — which can range from 0 to 10,000 — and the fact that noise tends to be proportional to signal strength. Using the log-transform of the data ensures that low level changes in expression receive relatively equal regard as those occurring at high levels. Therefore, our normalization routine begins by taking the log-transform of the raw expression data.

Our interest is in the relative changes in expression values, not absolute; so we centered each signal about 0 by subtracting its mean. That is, the $i^{th}$ gene expression signal, denoted by $\hat{G}_{i,*}$, once normalized will be

$$G_{i,*} = \ln \hat{G}_{i,*} - \frac{\sum_{j=1}^{T} \ln \hat{G}_{i,j}}{T},$$

(2.3)

where the log of a vector is the vector of the logarithms. The notation of the * symbol in the subscript will indicate the entirety of that index so that $\hat{G}_{i,*}$ is equivalent to $\hat{G}_{i,1:T}$.

### 2.3 Wavelets

A wavelet transform (WT) can be used to fit a linear model given a set of sample data points, $(y_1, \ldots, y_T)$. Instead of mapping the samples from a temporal domain into a frequency domain (as with Fourier transforms), the wavelet transform maps the samples into a time-frequency plane. A benefit of
a WT is that it can get as local as necessary, whereas the Fourier transform is
globalized over the entire signal. Wavelet transforms are capable of zooming in
or out at different locations of the signal. With wavelets, one can choose the
resolution level with which to transform a given signal [7].

The term \textit{wavelet} refers to a set of basis functions defined by a pair of
related functions: the wavelet generating function, denoted by \( \psi \), and the scaling
function, denoted by \( \phi \). These two functions work in concert to generate an
infinite set of basis functions. Typically, however, only a finite subset of these
basis functions is necessary to get a close approximation of \( f \). By choosing
different generating and scaling functions, many different families of wavelet
functions can be produced. In this thesis we focus on Haar wavelets, which are
specified as follows.

\subsection{The Haar wavelet}

For a function \( f(t), t \in [0,1] \), the Haar scaling function is defined as:

\[
\phi(t) = \begin{cases} 
1 & \text{if } 0 \leq t \leq 1; \\
0 & \text{otherwise},
\end{cases} \tag{2.4}
\]

and the Haar wavelet generating function is defined as:

\[
\psi(t) = \phi(2t) - \phi(2t - 1). \tag{2.5}
\]

or equivalently,

\[
\psi(t) = \begin{cases} 
1 & \text{if } 0 \leq t < \frac{1}{2}; \\
-1 & \text{if } \frac{1}{2} < t \leq 1; \\
0 & \text{otherwise},
\end{cases} \tag{2.6}
\]
The family of Haar basis functions is then defined by \( \{ \phi(t), \psi_{0,0}(t), \psi_{1,0}(t), \psi_{1,1}(t), \ldots \} \)
where \( \psi_{0,0}(t) = \psi(t) \) and
\[
\psi_{d,s}(t) = \psi(2^d t - s),
\]
for \( d \in \mathbb{Z}^+ \) and \( 0 \leq s \leq S_d = 2^d - 1 \).

Generalizing for a function \( f(t), t \in [a, b] \), the Haar scaling function is defined as:
\[
\phi(t) = \begin{cases} 
\frac{1}{\sqrt{b-a}} & \text{if } a \leq t \leq b; \\
0 & \text{otherwise},
\end{cases}
\]
and applying (2.5), the Haar wavelet generating function is defined as:
\[
\psi(t) = \begin{cases} 
\frac{1}{\sqrt{b-a}} & \text{if } a \leq t < \frac{a+b}{2}; \\
\frac{1}{\sqrt{b-a}} & \text{if } \frac{a+b}{2} < t \leq b; \\
0 & \text{otherwise}.
\end{cases}
\]
The denominator \( \sqrt{b-a} \) is chosen so that \( \int_a^b \phi(t)^2 = 1 \). Beginning with \( \phi(t) \)
and \( \psi(t) \), the remaining Haar basis functions on \([a, b]\) are then defined by
\[
\psi_{d,s}(t) = \frac{1}{2^\frac{d}{2}} \psi(a + 2^d \left( \frac{t-a}{b-a} \right) - s)(b-a)).
\]

2.3.2 Defining a finite basis using Haar wavelet functions

The orthonormal wavelet basis is generated by dilating and shifting \( \psi \) (see
Fig. 2.1). Think of \( d \) as indexing the dilation of \( \psi \), and \( s \) as indexing the shifting
of \( \psi \). By choosing a cutoff, \( D \), for \( d \), and thereby limiting how deeply to dilate \( \psi \), we can establish a finite basis consisting of \( 2^{D+1} \) functions and approximate
\( f \) by the expansion
\[
f(t) = c_0 \phi(t) + \sum_{d=0}^{D} \sum_{s=0}^{S_d} c_{d,s} \psi_{d,s}(t).
\]
For ease of notation, we re-index these functions with a single index, $k$, where $k = 1$ references the first basis function, $\psi_1(t) = \phi(t)$, and for $k = 2, \ldots, n = 2^{D+1}$, we define $\psi_k = \psi_{d,s}$, where $k = 2^d + s + 1$. With this re-indexing, (2.11) can be rewritten as

$$f(t) = \sum_{k=1}^{n} c_k \psi_k(t). \quad (2.12)$$

---

**Figure 2.1:** The eighth-order Haar wavelet basis functions on the interval $[0, 4]$. Here, $n = 8$ because $D = 2$.

---

### 2.3.3 Calculating the wavelet coefficients
The wavelet coefficients \(c_1, \ldots, c_n\) in (2.12) are chosen to minimize the least squares residual

\[
\sum_{j=1}^{T} (y_j - f(t_j))^2 = \sum_{j=1}^{T} \left( y_j - \sum_{k=1}^{n} c_k \psi_k(t_j) \right)^2 = \|y - cW\|^2,
\]

where \(y = (y_1, \ldots, y_T)\), \(c = (c_1, \ldots, c_n)\), and \(W\) is the \(n \times T\) design matrix defined by \(W_{kj} = \psi_k(t_j)\). The solution \(c\) is calculated by the equation

\[
c = yW^+,
\]

where \(W^+\) is the Moore-Penrose pseudoinverse of \(W\). \(W^+\) is calculated using the singular value decomposition (SVD) of \(W\):

\[
W = U \Sigma V^T,
\]

where the columns of \(U\) are the eigenvectors of \(WW^T\), the diagonal of \(\Sigma\) contains the singular values of \(W\), and the rows of \(V^T\) are the eigenvectors of \(W^TW\). \(W^+\) is defined by

\[
W^+ = V^T X^+ U,
\]

where

\[
X^+ = \begin{cases} 
0 & \text{if } X_{ii} = 0 \\
\frac{1}{X_{ii}} & \text{otherwise.}
\end{cases}
\]

When \(W\) has full rank and \(T \geq n\), then

\[
W^+ = (W^TW)^{-1}W^T.
\]

When \(W\) has full rank and \(T < n\), then

\[
W^+ = W^T(WW^T)^{-1}.
\]
In general, (2.18) or (2.19) makes possible the least-squares solution to (2.14).

We define the $m \times n$ matrix $C$ as the set of all wavelet coefficients resulting from the dataset $G$, where $C_{i,k}$ is the $k^{th}$ wavelet coefficient for the $i^{th}$ gene. $C$ is calculated by the equation

$$C = GW^+. \quad (2.20)$$

### 2.3.4 Haar wavelet coefficients

The first of the wavelet coefficients, $C_1$, is referred to as the sparse coefficient and is the average value over all $t_j$, that is $C_1 = \frac{\sum_j f(t_j)}{T}$. Because of our normalization scheme, the calculated value of $C_1$ will always be 0. The remaining wavelet coefficients, $C_2, ..., C_n$, are referred to as the detail coefficients. Each detail coefficient corresponds to a localized frequency band, or scale, of the signal.

### 2.4 Identification of Significant Signals

Because the relative magnitude of the Haar wavelet coefficients was the key factor in our analysis of significant temporal patterns, it was necessary to gain a sense of how the coefficients were distributed. We used bootstrapping to estimate the background distribution of each gene’s associated wavelet coefficients, thereby obtaining a $p$-value for each coefficient $C_k$, $k \geq 2$. Details of this are provided in Section 2.4.1.

Using the smallest $p$-values from each signal, we calculate the probabilities of their smallness based upon their overall cumulative distribution; essentially, calculating a $p$-value for each minimum $p$-value. Applying a FDR procedure to these $p$-values, we determine the cutoff point at which to reject the hypothesis that a signal exhibits no significant temporal pattern.
2.4.1 Bootstrapping procedure

In our method, we wanted to identify genes with significantly large wavelet coefficients. But we needed to find a way to distinguish significance. This was accomplished using a bootstrapping procedure to estimate a p-value for the magnitude of each wavelet coefficient $|C_{i,k}|$.

For each gene $i$ we generated a set of $H$ trial genes $\tilde{G}_{i,h}$, for $h = 1, \ldots, H$. Each trial gene is generated by sampling (with replacement) from expression values $(\tilde{G}_{i,1}, \ldots, \tilde{G}_{i,T})$, and then normalizing the resulting expression profile according to (2.3). The Haar wavelet coefficients $\tilde{C}_{i,*h}$ were then calculated for each trial gene according to (2.14).

The trial coefficients calculated above enable us to calculate a p-value for the magnitude of each wavelet coefficient. To do this, for each $(i,k)$, we calculate the fraction of corresponding coefficients $\tilde{C}_{i,k,h}$ for which

$$|\tilde{C}_{i,k,h}| \geq |C_{i,k}|.$$  \hspace{1cm} (2.21)

More precisely, let $P_{i,k}$ denote this estimated p-value for the $k^{th}$ coefficient for gene $i$. $P_{i,k}$ is calculated by the equation

$$P_{i,k} = \frac{\tilde{h} + 1}{H},$$ \hspace{1cm} (2.22)

where $\tilde{h}$ is the number of trials $h$ for which (2.21) is satisfied.

Note that if the data were completely random, the p-values should reflect a uniform distribution. Moreover, because the Haar wavelet functions are orthogonal, the p-values are also independent. These two facts enable us to calculate a distribution for the smallest p-value.
The final step begins by selecting the smallest $p$-value associated with gene $i$, which we notate as $P_{i,k_{\text{min}}}$. Since the $k^{\text{th}}$ set of coefficients $(C_{*,k})$ is independent and uniformly distributed, we can treat all $P_{i,k}$ as random variables and determine the likelihood that $P_{i,k_{\text{min}}}$ is relatively small. Essentially, the final step is to calculate a $p$-value for each $P_{i,k_{\text{min}}}$. To do this, we use the random variables $P_{i,2:n}$ associated with each gene $i$ to calculate the probability that the smallest of $n - 1$ independent and uniformly distributed random variables is at least as small as $P_{i,k_{\text{min}}}$. This probability is calculated by:

$$P_i = \text{Prob}(\min(x_2, \ldots, x_n) \leq P_{i,k_{\text{min}}}) = 1 - (1 - P_{i,k_{\text{min}}})^{n-1}$$

(2.23)

where random variables $x_k \sim \text{Uniform}(0,1)$. It is this set of probabilities to which we apply the false discovery rate procedure.

### 2.4.2 FDR analysis

In multiple testing, the number of type I errors (false rejections) can increase considerably. For example, when testing 10,000 genes using a significance level of $\alpha = 0.05$, we would expect to identify approximately 500 false discoveries ($FD$) by chance alone. If our multiple tests resulted in 1000 discoveries ($D$), then the expected rate of false discoveries would be $\frac{FD}{D} = \frac{500}{1000} = 0.5$, which is rather large. Therefore, instead of controlling the chance of a single false discovery occurring (e.g., via Bonferroni methods), we use a false discovery rate (FDR) procedure developed by Benjamini and Hochberg [6] to control the overall proportion of false discoveries.

Given a pre-defined significance level, $\alpha$, and a set of sorted $p$-values $P$ to be compared, the FDR procedure defines a multiple testing threshold we will
call $FDT$, such that

$$FDT = \max \{ P_i : \; P_i \leq \frac{i}{m} \}.$$  \hfill (2.24)

So instead of using a static significance level as in the Bonferroni method of multiple testing, we implement a dynamic threshold which is a function of the static significance level ($\alpha$) and the particular set of $p$-values being assessed. We are essentially calculating a significance level for each $i$ as follows:

$$\alpha_i = \frac{i}{m} \alpha.$$  \hfill (2.25)

By defining $\alpha_i$ in this way, if $P_i < \alpha_i$ then the desired FDR is achieved. One can simply scan down $P$ in descending order, until the first index $i$ for which $P_i < \alpha_i$ is reached, and then reject the null hypothesis ($H_0 :$ there is no significant temporal pattern) for all $j \geq i$.

This FDR procedure ensures that the expected proportion of false discoveries ($\mathbb{E}(\frac{FP}{P})$) as a function of $FDT$, will not exceed the predefined significance level $\alpha$ over multiple hypothesis tests. For example, if 100 out of 1000 genes are found to be temporally significant, and $\alpha$ was chosen to be 0.2, then on average we can expect at least 80 of these to be valid discoveries. Our algorithm was tested using a predefined significance level of $\alpha = 0.1$. Since we knew that only a small proportion of $P$ would be rejected under the null hypothesis, $P$ was sorted and searched in ascending order for the sake of efficiency. The pseudocode in Tab. 2.1 summarizes our method.

2.5 Algorithm Summary and Implementation

The complete algorithm is summarized by the pseudocode in Tab. 2.2. Input parameters are listed in Tab. 2.3. Algorithm output was in the form of
Table 2.1: FDR procedure pseudocode

Step 0: (Initialization) Given a vector of $m$ p-values, $P$, and a predefined significance level $\alpha$;

Step 1: (Sort) Sort $P$ in ascending order;

Step 2: (Critical Significance Levels) For each $i$, calculate $\alpha_i$ according to (2.25);

Step 3: (Find Threshold) For each $i$, compare $P_i$ to $\alpha_i$; when $P_i > \alpha_i$, reject the null hypothesis for all $\hat{i} < i$;

Step 4: (Desert and Return Decision Vector) Return a decision vector of length $m$ in the original order of $P$, which indicates a discovery (1) or a nondiscovery (0) for each gene $i$.

A decision vector, $z$, of length $m$ where $z_i = 0$ indicated that the $i^{th}$ signal was not flagged as significant, and $z_i = 1$ indicated otherwise.
Table 2.1: FDR procedure pseudocode

<table>
<thead>
<tr>
<th>Step 0 : (Initialization)</th>
<th>Given a vector of $m$ p-values, $P$, and a predefined significance level $\alpha$;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1 : (Sort)</td>
<td>Sort $P$ in ascending order;</td>
</tr>
<tr>
<td>Step 2 : (Critical Significance Levels)</td>
<td>For each $i$, calculate $\alpha_i$ according to (2.25);</td>
</tr>
<tr>
<td>Step 3 : (Find Threshold)</td>
<td>For each $i$, compare $P_i$ to $\alpha_i$; when $P_i &gt; \alpha_i$, reject the null hypothesis for all $i' &lt; i$;</td>
</tr>
<tr>
<td>Step 4 : (Desert and Return Decision Vector)</td>
<td>Return a decision vector of length $m$ in the original order of $P$, which indicates a discovery (1) or a nondiscovery (0) for each gene $i$.</td>
</tr>
</tbody>
</table>

A decision vector, $z$, of length $m$ where $z_i = 0$ indicated that the $i^{th}$ signal was not flagged as significant, and $z_i = 1$ indicated otherwise.
<table>
<thead>
<tr>
<th>Step 0: (Initialization)</th>
<th>Given an $m \times T$ matrix of expression values $G$, a vector of $T$ time points $t$, input parameters $a$, $b$, $R$, $H$, and $\alpha$ (see description of input parameters in Fig. 2.3);</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1: (Normalization)</td>
<td>Construct the normalized dataset, $G$, according to formula 2.3;</td>
</tr>
<tr>
<td>Step 2: (Haar Wavelet Decomposition)</td>
<td>Generate the wavelet coefficients, $C$, according to formula 2.14;</td>
</tr>
<tr>
<td>Step 3: (Estimation of $P$-values)</td>
<td>As described in the Section 2.4.1, use $H$ to estimate the $p$-value for each wavelet coefficient using formula 2.22, then calculate $P$, for each gene using formula 2.23;</td>
</tr>
<tr>
<td>Step 4: (FDR Analysis)</td>
<td>As described in Fig. 2.1, use $P$ and $\alpha$ to determine temporal significance for each gene.</td>
</tr>
</tbody>
</table>
**Table 2.3:** Algorithm input parameters

- **t**, a vector of length $T$, giving the ordered time points at which expression levels were sampled;

- **a** and **b**, endpoints indicating the starting and ending times, used to define the Haar wavelet functions;

- **R**, an integer $> 0$, specifying the size of the Haar basis, $n = 2^R$;

- **G**, an $m \times T$ matrix containing the expression values of $m$ signals, each sampled at $T$ time points indexed by $t_j$;

- **H**, an integer $> 0$, specifying the number of trial signals to generate for each gene in the bootstrapping procedure; and

- **α**, a value in the interval (0,1), stipulating the false discovery rate to be used in deciding the temporal significance of genes.
3. TESTING AND RESULTS

The method was tested using simulated signals representing the expression levels across time for a set of genes, \( \hat{G} \). Some of the gene signals were generated only by random noise, and the rest by a predefined "true" temporal pattern \( f \) plus random noise. To generate temporal patterns for the "true" signals, we used three families: the Haar step function, the sine function, and non-cyclical regression functions (exponential and polynomial). The added noise was Gaussian with mean 0 and with standard deviation, \( \sigma \), varied to reflect different noise-to-signal ratios. Finally, the algorithm was applied to the genome-wide expression levels measured from the mammary tissue of rattus norvegicus across the four day estrous cycle.

3.1 Generating Simulation Time Series

In order to observe our algorithm's performance at both detecting the "true" signals and dismissing the "false", we generated a variety of simulated datasets used to represent \( \hat{G} \) in calling our primary algorithm. Each simulated dataset was made up of two kinds of signals – those originating only from random noise, and those produced from a time-dependent generating function \( f \) with Gaussian noise added. The latter kind we refer to as the "true" signals. Since we knew which was which, we were able to calculate actual false discovery and sensitivity rates of our algorithm's performance.

For each \( f \) tested, a set of simulated signals, \( F \), was generated so that a specifiable fraction \( p \), of them was based upon \( f + \text{noise} \), and the rest consisted
only of noise. The matrix, \( F \), consisting of \( m \) simulated signals with \( T \) time points (placed according to \( t \_rate \)) was constructed such that \( p \) of them were "true", and defined as \( F_i = f + N(0, \sigma^2) \), while the remaining were designed to be pattern-free, defined as \( F_i = 0 + N(0, \sigma^2) \). Sampling rate, sampling frequency, and degree of noise-to-signal ratio were varied during testing in order to analyze their effects upon our algorithm's identification of temporal patterns. The testing scheme utilized the parameters listed in Fig. 3.1.

For simplicity and ease of comparability, all simulations arranged the time points \( t \) within the interval \([0, 1]\) (that is, \( a = 0, b = 1 \)), and we used \( R = 3 \) (see Fig. 2.3), thereby keeping the number of basis functions at 8. Furthermore, since each \( f \) used for simulation had a variance of less than 5, \( F \) was assumed to be already log-transformed once sent to the main algorithm. Therefore, the normalization of \( F \) consisted only of centering each \( F_i \) about 0.

\( F \) was analyzed by our primary algorithm using \( H = 1000 \) and \( \alpha = 0.1 \) to test for significantly large coefficients. We counted the resulting number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). These variables were then used to calculate the observed FDRs: \( FDR_{obs} = \frac{FP}{FP+TP} \); and the observed sensitivity rates: \( SR_{obs} = \frac{TP}{TP+FN} \).

For each generating function \( f \), we produced simulations over various combinations of \( \eta, T, p, \) and \( t \_rate \). The testing parameters varied as follows:

- \( 0.1 \leq \eta \leq 2.0 \);
- \( T = [6, 7, 8, 9, 10, 15, 20, 25, 30, 40, 50] \);
- \( p = [0.1, 0.25, 0.5, 0.75, 0.9] \).
Table 3.1: Input parameters needed to construct $F$

- $f$, a function to be used in generating the “true” signals in $F$;
- $\eta$, a value used to determine the noise-to-signal ratio such that the noise ($N(0, \sigma^2)$) added to each signal, $F_i$, had standard deviation $\sigma = \eta \cdot f_{max}$, where $f_{max}$ is the maximum amplitude of zero-centered $f$.
- $T$, an integer indicating the number of time points with which to populate $t$;
- $t\_rate$, a string indicating the type of sampling interval with which to build $t$; either a uniform (“even”) or random (“rand”) placement of $T$ time points;
- $m$, an integer dictating the number of simulated signals to generate;
- $p$, a value in the interval $[0,1]$ indicating the fraction of $m$ signals to generate using $f$; the fraction of “true” signals populating $F$;
- $H$, an integer $> 0$, specifying the number of trial signals to generate for each signal in the bootstrapping procedure; and
- $\alpha$, a value in the interval $(0,1)$, stipulating the false discovery rate used in determining the significance of wavelet coefficients.

$FDR_{obs}$ and $SR_{obs}$ were recorded and stored for each scenario. Output was generated in tabular form of $\eta$ against $T$ (see Tab. 3.2). For each value of $p$, we
generated four $\eta$-$T$ tables: $FDR_{\text{obs}}$ from evenly- and unevenly-spaced $T$, and $SR_{\text{obs}}$ for evenly- and unevenly-spaced $T$. Each table was the average result over 10 simulations for a given $f$, $\eta$, $T$, $p$, and $t_{rate}$.

By varying the noise-to-signal ratio ($\eta$), the number of sample points ($T$), the sampling rate ($t_{rate}$) and the proportion ($p$) of $f$-generated signals, we were able to compare results of various scenarios within each type of generating function. The motivation for each type of test is to get a sense of how many time points are needed to identify a significant signal, and how their placement (uniform or not) and the degree of added noise affects this number.

3.1.1 Haar wavelet signals

The first set of tests involved signals based on the Haar wavelet itself. We used the step function

$$\psi(t) = \begin{cases} 
1 & \text{if } 0 \leq t < \frac{1}{2}; \\
-1 & \text{if } \frac{1}{2} < t \leq 1,
\end{cases}$$

to act as $f$. These signals clearly display a low frequency pattern across the

![Figure 3.1: The Haar Wavelet step function, $\psi$](image)

28
overall time interval, i.e., high value to low value. The motivation behind implementing this test was to see how effective the algorithm would be on the simplest possible test function. This test served as a good reference base with which to compare the other, more complex tests.

3.1.2 Sinusoidal signals

The second set of tests used variations of the sine function with which to build $F$. This test was based on the standard sine function

$$f(t) = \sin(\text{freq} \cdot \pi \cdot t),$$

where the parameter \text{freq} was used to vary the frequency from approximately half of a period to approximately two periods on the interval $t = [0, 1]$. We varied \text{freq} within the interval $[0.05, 16]$. See Fig. 3.2 for some of the frequencies we tested. These are signals which display windows of cyclical patterns. In varying

![Sinusoidal functions of various frequencies on the time interval $[0,1]$.](image)

**Figure 3.2:** Sinusoidal functions of various frequencies on the time interval $[0,1]$. 

29
the frequency we intended to test whether the algorithm favors lower or higher frequencies.

3.1.3 Regression signals

The last set of tests involved regression functions that have been previously used in various Monte Carlo performance studies [25]. This set consists of the following four functions:

\[
\begin{align*}
    f_{\text{regr}1}(t) &= 1 - 48t + 218t^2 - 315t^3 + 145t^4; \\
    f_{\text{regr}2}(t) &= 0.3e^{-64(t-0.25)^2} + 0.7e^{-256(t-0.75)^2}; \\
    f_{\text{regr}3}(t) &= 10e^{-10t}; \\
    f_{\text{regr}4}(t) &= \begin{cases} 
    e^{(t-\frac{1}{3})}, & \text{for } t < \frac{1}{3}; \\
    e^{-2(t-\frac{1}{3})}, & \text{for } t \geq \frac{1}{3}.
    \end{cases}
\end{align*}
\]

These generating functions are neither simple — as with the Haar step function,

![Graphs of regression functions](image)

(a) $f_{\text{regr}1}(t)$  (b) $f_{\text{regr}2}(t)$  (c) $f_{\text{regr}3}(t)$  (d) $f_{\text{regr}4}(t)$

**Figure 3.3:** Regression functions 1 through 4, on the interval $[0,1]$.

nor cyclic — as with the sinusoidal functions. These generating functions are implemented in an effort to see how the algorithm fares when dealing with more
complex signals. The first regression function, Fig. 3.3(a), displays a trend with a little fine structure while the third, Fig. 3.3(c), displays trend but with no fine structure. The second, Fig. 3.3(b), has obvious time dependent variation in curvature and the fourth, Fig. 3.3(d), is a function whose first derivative is undefined at \( t = \frac{1}{3} \), contradicting standard assumptions for optimal performance. Each regression function was used to generate “true” signals such that \( \hat{f}(t) = f(t) + N(0, \sigma^2) \).

3.2 Results: Simulation Testing

Each simulation generated \( m = 100 \) signals, and used \( H = 1000 \) and \( \alpha = 0.1 \) in determining significant temporal activity. Results were averaged over 10 iterations.

3.2.1 Performance using Haar-based \( f \)

The Haar step function simulations yielded similar results for evenly and randomly spaced \( T \) (see Tables 3.2, 3.3 and Fig. 3.4). With 10% of signals true \( (p = 0.1) \), the observed sensitivity rates stayed above 90% when \( T \) was as small as 15 and \( \eta \leq 0.5 \), and when \( \eta \) was as large as 1.0 and \( T \geq 40 \).

Table 3.2: Observed sensitivity rates for \( f = \psi(t) \), evenly spaced \( T \) and \( p = 0.1 \)

<table>
<thead>
<tr>
<th>( T )</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.07</td>
<td>0.09</td>
<td>0.12</td>
<td>0.24</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.03</td>
<td>0.09</td>
<td>0.21</td>
<td>0.29</td>
<td>0.73</td>
<td>0.89</td>
</tr>
<tr>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.33</td>
<td>0.66</td>
<td>0.93</td>
<td>0.99</td>
</tr>
<tr>
<td>0.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
<td>0.23</td>
<td>0.87</td>
<td>0.98</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0.80</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0</td>
<td>0.01</td>
<td>0.11</td>
<td>0.43</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0.03</td>
<td>0.03</td>
<td>0.10</td>
<td>0.53</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Table 3.3: Observed sensitivity rates for \( f = \psi(t) \), randomly spaced \( T \) and \( p = 0.1 \)

<table>
<thead>
<tr>
<th>( \eta )</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.07</td>
<td>0.07</td>
<td>0.11</td>
<td>0.14</td>
<td>0.31</td>
</tr>
<tr>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.03</td>
<td>0.03</td>
<td>0.17</td>
<td>0.19</td>
<td>0.35</td>
</tr>
<tr>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.08</td>
<td>0.30</td>
<td>0.76</td>
<td>0.95</td>
<td>0.99</td>
</tr>
<tr>
<td>0.75</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.07</td>
<td>0.19</td>
<td>0.56</td>
<td>0.79</td>
<td>0.96</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0.06</td>
<td>0.61</td>
<td>0.90</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.15</td>
<td>0.30</td>
<td>1.00</td>
<td>0.95</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.17</td>
<td>0.22</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

(a) Evenly spaced \( T \)  
(b) Randomly spaced \( T \)

Figure 3.4: Surface plots of observed sensitivity rates. (These are 3-dimensional renderings of Tables 3.2 and 3.3, viewed from their upper left hand corners towards their lower right.)

When added noise was at its best (smallest; \( \eta = 0.1 \)), \( T \) could be as small as 15 and still identify 100% of the true signals, with \( FDR_{\text{obs}} = 0.03 \) for \( T \) evenly spaced and \( FDR_{\text{obs}} = 0.06 \) for \( T \) randomly spaced. When \( T \) dropped from 15 to 10, sensitivity decreased to 53% (\( FDR_{\text{obs}} = 0.02 \)) for evenly spaced, and to 22% (\( FDR_{\text{obs}} = 0.04 \)) for randomly spaced time points.

To get a better idea of what was going on between \( T = 15 \) and \( T = 10 \), we reran the simulations using the \( \eta = 0.1 \) and \( 9 \leq T \leq 16 \). Fig. 3.5 shows the
Figure 3.5: Sensitivity rates for $16 \geq T \geq 9$ with $\eta = 0.1$.

resulting sensitivity rates for evenly and randomly spaced $T$. We can see that, while in Fig. 3.4 there appears to be a big drop between $T = 15$ and $T = 10$, Fig. 3.5 shows that the decrease is actually gradual, except at $T = 12$, where there occurs an increase for evenly spaced $T$ and a decrease for randomly spaced $T$.

When the number of time points was at its best (largest; $T = 50$), a relatively large amount of added noise ($\eta = 1.5$) did not prevent the algorithm from capturing 89% ($FDR_{obs} = 0.14$) and 76% ($FDR_{obs} = 0.14$) of the true signals, for evenly and randomly spaced $T$, respectively. However, when $\eta$ was increased to 2.0, even a large sample size could not make up for the damaging effect of too much noise; the algorithm identified 46% ($FDR_{obs} = 0.1$) and 31% ($FDR_{obs} = 0.09$) of the true signals, for evenly and randomly spaced $T$, respectively. Fig. 3.6 shows how when $\eta = 1.5$ the signals originating from the Haar step function are still visible, but when $\eta = 2.0$ they are not as easily distinguished from the
Figure 3.6: Haar wavelet-generated (normalized) signals for $T = 40$, where $T$ is evenly spaced and $p = 0.1$. This illustrates how, even with a large number of uniform samples, true signals (black) can quickly become nearly indistinguishable from pure noise (grey) when $\eta$ increases from 1.5 to 2.0.

For both evenly and randomly spaced time points, $FDR_{obs}$ remained below 0.2 for $\eta \leq 1.0$ and $T \geq 20$. Unlike the fluid patterns seen with the observed sensitivity rates, $FDR_{obs}$ were generally more erratic – with surface plots more reminiscent of origami than landscapes (see Fig. 3.7). For $T < 10$, the typical outcome was that no false positives occurred; this was often a by-product of there being no positives tagged at all. Positives (true or not) did not occur at all when $T$ was smaller than 7.

There were little to no differences seen in results as the proportion of true signals ($p$) was increased. Observed FDRs approached zero as $p$ increased, but the sensitivity rates were largely unchanged (see Fig. 3.8).
Figure 3.7: Surface plots for the observed FDRs from uniform and randomly spaced $T$, $p = 0.1$. 

(a) Evenly spaced $T$  

(b) Randomly spaced $T$
Figure 3.8: Surface plots of observed sensitivity when the proportion of true signals (p) is 50\% (a and b) and 90\% (c and d). No significant differences were seen between small and large p.
3.2.2 Performance using sinusoidal-based $f$

The results from the sinusoidal simulations were similar to those generated using the Haar step function in that as $p$ increased from 0.1 to 0.9, there were only small improvements observed. Therefore, we will present only those results associated with $p = 0.1$. Overall, lower frequencies (frq < 5) yielded better results than higher frequencies (frq $\geq 5$). This contrast is illustrated in Figures 3.10 and 3.12. Haar wavelet functions are aligned with sinusoidal frequencies of a power of two, so there were bursts of improvement near these frequencies, with deterioration in between. This trend can be seen in Fig. 3.11 for frequencies near 2 and 4, and in Fig. 3.12 for frequency of 8. Fig. 3.9 plots the mean sensitivity rates (for evenly spaced $T$, $\eta = 0.1$) as a function of frequency. Our algorithm performed optimally for frequencies near 2 (see Fig. 3.11).

![Figure 3.9: Mean sensitivity rate across all evenly spaced $T$ (with $\eta = 0.1$) as a function of frequency rate.](image)

37
Figure 3.10: Surface plots of the resulting sensitivity from low frequency (near 0 and 1) sinusoidal-generated simulations.
Figure 3.11: Surface plots of the resulting sensitivity from lower frequency (near 2 to 4) sinusoidal-generated simulations.
Figure 3.12: Surface plots of the resulting sensitivity from higher frequency (greater than 5) sinusoidal-generated simulations.
Generally, for all frequencies examined, results improved when $T$ was evenly spaced. However, when $T$ was small, the algorithm often performed slightly better for randomly spaced $T$. Here we begin our discussion of the results from some of the lower frequencies (featured in Figures 3.10 and 3.11).

**Table 3.4:** Observed sensitivity rates for $f = sin(0.3\pi t)$, $T$ evenly spaced and $p = 0.1$

<table>
<thead>
<tr>
<th>$T$</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.05</td>
<td>0.06</td>
<td>0.18</td>
<td>0.35</td>
</tr>
<tr>
<td>0.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0.09</td>
<td>0.18</td>
<td>0.18</td>
<td>0.40</td>
<td>0.81</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.17</td>
<td>0.64</td>
<td>0.73</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.14</td>
<td>0.78</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.17</td>
<td>0.62</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

When frequency was set to 0.3 (see Tab. 3.4 and Fig. 3.10 e and f), minimal added noise ($\eta = 0.1$) required as little as 20 evenly spaced sample points to identify 92% ($FDR_{obs} = 0.07$) of the true signals, or 25 randomly spaced sample points to identify 90% ($FDR_{obs} = 0.07$). However, if evenly spaced $T$ dropped from 20 to 15, the sensitivity dove from 92% to 17% ($FDR_{obs} = 0.15$). The decline was more gradual for randomly spaced $T$.

Even for large $T$, added noise took a toll on sensitivity rates. For evenly spaced $T = 50$, sensitivity rates dropped from 81% ($FDR_{obs} = 0.16$) to 35% ($FDR_{obs} = 0.15$) as $\eta$ increased from 0.75 to 1.0. For randomly spaced $T = 50$, sensitivity dropped from 97% ($FDR_{obs} = 0.09$) to 48% ($FDR_{obs} = 0.08$) as $\eta$ increased from only 0.5 to 0.75.

To get a better idea of what was going on in between our standard simulation values of $T$ and $\eta$, we ran additional simulations using small $\eta$ with $T$ in finer
detail, and large $T$ with $\eta$ in finer detail. We can see from Fig. 3.13, which shows how $T$ (with minimal added noise) affects sensitivity, that the largest jump actually occurs between evenly spaced $T = 16$ and $T = 15$. From Fig. 3.14, which shows how $\eta$ (with large $T$) affects sensitivity, we see that the largest change for both evenly and randomly spaced $T$ actually occurs between $\eta = 0.7$ and $\eta = 0.9$.

![Graph showing sensitivity rates for frq = 0.3, $\eta = 0.1$, as a function of $T$ in finer detail. Sensitivity decreases more sharply for evenly spaced $T$ (solid dark line) between 20 and 15 than for randomly spaced $T$ (dashed light line).]

**Figure 3.13:** Sensitivity rates for frq = 0.3, $\eta = 0.1$, as a function of $T$ in finer detail. Sensitivity decreases more sharply for evenly spaced $T$ (solid dark line) between 20 and 15 than for randomly spaced $T$ (dashed light line).
Figure 3.14: Sensitivity rates for $frq = 0.3$, $T = 50$, as a function of $\eta$ in finer detail. Sensitivity decreases most when $0.5 \leq \eta \leq 1.0$. Evenly (solid dark line) and randomly (dashed light line) spaced large $T$ produce similarly decreasing rates.
Table 3.5: Observed sensitivity rates for $f = \sin(1.3\pi t)$, evenly spaced $T$ and $p = 0.1$

<table>
<thead>
<tr>
<th>$\eta$</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.01</td>
<td>0.10</td>
</tr>
<tr>
<td>0.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.05</td>
<td>0.02</td>
<td>0.06</td>
<td>0.06</td>
<td>0.23</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0.02</td>
<td>0.06</td>
<td>0.31</td>
<td>0.57</td>
</tr>
<tr>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0.10</td>
<td>0.37</td>
<td>0.92</td>
<td>0.99</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.18</td>
<td>0.76</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Sensitivity rates for frequencies near 1 were not as high as those for frequencies near 0. For frequency 1.3 (see Tab. 3.5 and Fig. 3.10 k and l), a minimal amount of noise still required an evenly spaced $T$ as large as 30 to be able to identify 76% ($FDR_{obs} = 0.14$) of the true signals. If the 30 time points were randomly spaced, only 38% ($FDR_{obs} = 0.17$) were identified. And if evenly spaced $T$ dropped from 30 to 25, the sensitivity rate dropped from 76% to 18% ($FDR_{obs} = 0.14$).

We again ran additional simulations using small $\eta$ with $T$ in finer detail, and large $T$ with $\eta$ in finer detail. Fig. 3.15, which plots sensitivity as a function of $T$ (with minimal added noise), shows that evenly spaced $T$ is adversely affected for $T < 26$, whereas randomly spaced $T$ has a more gradual affect on sensitivity. Sensitivity rates for large $T$ behaved similarly for evenly and randomly spaced $T$ as the level of noise changed. We can see from Fig. 3.16, which shows how $\eta$ (with large $T$) affects sensitivity, that for both evenly and randomly spaced $T$, rates become quite low as $\eta$ approaches 0.5.
Figure 3.15: Sensitivity rates for $frq = 1.3$, $\eta = 0.1$, as a function of $T$ in finer detail. Sensitivity decreases more sharply for evenly spaced $T$ (solid dark line) between 30 and 25 than for randomly spaced $T$ (dashed light line).

Figure 3.16: Sensitivity rates for $frq = 1.3$, $T = 50$, as a function of $\eta$ in finer detail. Evenly (solid dark line) and randomly (dashed light line) spaced large $T$ produce similarly decreasing rates.
Table 3.6: Observed sensitivity rates for \( f = \sin(1.8\pi t) \), evenly spaced \( T \) and \( \eta = 0.1 \)

<table>
<thead>
<tr>
<th>( \eta )</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0.06</td>
<td>0.02</td>
<td>0.36</td>
<td>0.02</td>
</tr>
<tr>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.02</td>
<td>0.99</td>
<td>0.17</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
<td>0.11</td>
<td>0.12</td>
<td>0.22</td>
<td>0.50</td>
<td>0.72</td>
</tr>
<tr>
<td>0.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.24</td>
<td>0.48</td>
<td>0.58</td>
<td>0.95</td>
<td>0.99</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.07</td>
<td>0.67</td>
<td>0.88</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0</td>
<td>0.48</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

We’ve seen that the Haar wavelet transform favors low frequencies, and frequencies near powers of two. So it is not surprising that our algorithm performed optimally for frequencies near 2, the smallest power of two. When frequency was set to 1.8, (Tab. 3.6 and Fig. 3.11 b and c) even for relatively small \( T \) and large \( \eta \), sensitivity rates remained fairly high. For \( T = 20 \) and \( \eta = 0.5 \), the sensitivity rate was still as high as 67% (\( FDR_{obs} = 0.07 \)).

When added noise was most minimal (\( \eta = 0.1 \)), evenly spaced \( T \) could be as small as 15 and our algorithm was still able to identify virtually all temporal signals (\( FDR_{obs} = 0.07 \)); for randomly spaced \( T = 15 \), our algorithm identified 73% of temporal signals (\( FDR_{obs} = 0.04 \)). However, if evenly spaced \( T \) dropped from 15 to 10, the sensitivity rates fell from 99% to 0% (with \( FDR_{obs} \) undefined as no positives were tagged at all).

We ran additional simulations using small \( \eta \) with \( T \) in finer detail, and large \( T \) with \( \eta \) in finer detail. We can see from Fig. 3.17, which shows how \( T \) (with minimal added noise) affects sensitivity, that the largest jump actually occurs between evenly spaced \( T = 13 \) and \( T = 12 \). Fig. 3.18 shows that large \( T \) was able to withstand relatively high levels of added noise whether \( T \) was evenly or
randomly spaced.

![Graph](image1)

**Figure 3.17:** Sensitivity rates for $frq = 1.8$, $\eta = 0.1$, as a function of $T$ in finer detail. Sensitivity decreases more sharply for evenly spaced $T$ (solid dark line) between 14 and 12 (see arrows) than for randomly spaced $T$ (dashed light line).

![Graph](image2)

**Figure 3.18:** Sensitivity rates for $frq = 1.8$, $T = 50$, as a function of $\eta$ in finer detail. Evenly (solid dark line) and randomly (dashed light line) spaced large $T$ produce similarly decreasing rates and both have a high tolerance to a large amount of added noise.
Table 3.7: Observed sensitivity rates for $f = \sin(2.5\pi t)$, evenly spaced $T$ and $p = 0.1$

<table>
<thead>
<tr>
<th>$\eta$</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0.01</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
<td>0</td>
<td>0.04</td>
<td>0.07</td>
<td>0.21</td>
</tr>
<tr>
<td>0.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.04</td>
<td>0.06</td>
<td>0.18</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.07</td>
<td>0.22</td>
<td>0.50</td>
</tr>
<tr>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.01</td>
<td>0.39</td>
<td>0.58</td>
<td>0.92</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.77</td>
<td>0.56</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

For frequency 2.5 (see Tab. 3.7 and Fig. 3.11 h and i), results fared slightly better than when frequencies were near 1, but were worse than when frequencies were near 0. Minimal added noise ($\eta = 0.1$) needed evenly spaced $T = 25$ to identify 77% ($FDR_{obs} = 0.08$), but if $T = 30$, sensitivity rates curiously drop to 56% ($FDR_{obs} = 0.14$). This same is seen for randomly spaced $T$ where sensitivity is 64% ($FDR_{obs} = 0.14$) when $T = 25$, but only 49% ($FDR_{obs} = 0.12$) when $T = 30$. Additionally, if evenly spaced $T$ dropped from 25 to 20, the sensitivity rate went from 77% to 2% ($FDR_{obs} = 0.6$).

This pattern can be seen in Fig. 3.19. Note that the results appearing in Fig. 3.19 differ from those in Tab. 3.7 and Fig. 3.11 because they were obtained from a separate batch of simulations investigating detailed $\eta$ and $T$. The same patterns still emerge however.

For large $T$, $\eta$ affected sensitivity similarly for evenly and randomly spaced time points, except when $\eta = 0.6$ (see Fig. 3.20).
Figure 3.19: Sensitivity rates for $\text{freq} = 2.5$, $\eta = 0.1$, as a function of $T$ in finer detail. Sensitivity for evenly spaced $T$ (solid dark line) is larger when $T = 24$ than when $T = 29$, and a large decrease occurs between $T = 24$ and 23. For randomly spaced $T$ (dashed light line), sensitivity is larger when $T = 22$ than when $T = 31$.

Figure 3.20: Sensitivity rates for $\text{freq} = 2.5$, $T = 50$, as a function of $\eta$ in finer detail. Evenly (solid dark line) and randomly (dashed light line) spaced large $T$ produce similarly decreasing rates except for $0.5 \leq \eta \leq 0.7$, where randomly spaced $T$ yields smaller sensitivity rates.
Table 3.8: Observed sensitivity rates for $f = \sin(3.8\pi t)$, evenly spaced $T$ and $p = 0.1$

<table>
<thead>
<tr>
<th>$\eta$</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.01</td>
<td>0.04</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0.07</td>
<td>0.15</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>0.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.08</td>
<td>0.15</td>
<td>0.27</td>
<td>0.63</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.05</td>
<td>0.28</td>
<td>0.88</td>
<td>0.99</td>
</tr>
<tr>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.50</td>
<td>0.86</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
<td>0.77</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

When frequency was set to 3.8 (see Tab. 3.8 and Fig. 3.11 n and o), results were an improvement on a frequency of 2.5, but not as good as when frequencies were nearer to 2. A minimal level of added noise ($\eta = 0.1$) required 25 evenly spaced sample points in order to identify 77% ($FDR_{obs} = 0.08$) of the true signals, or 25 randomly spaced sample points to identify 66% ($FDR_{obs} = 0.10$). However, if evenly spaced $T$ dropped from 25 to 20, the sensitivity went from 77% to 5% ($FDR_{obs} = 0.44$). The decrease in sensitivity rates for randomly spaced $T$ however appeared more gradual. In Fig. 3.21, we can see that while there is a big difference in sensitivity rates between evenly spaced $T = 25$ and 20, the decrease is actually relatively uniform. For large $T$, $\eta$ affected sensitivity similarly for evenly and randomly spaced time points (see Fig. 3.22).
Figure 3.21: Sensitivity rates for freq = 3.8, \( \eta = 0.1 \), as a function of \( T \) in finer detail. Sensitivity drops significantly, but uniformly, for evenly spaced \( T \) (solid dark line) between 24 and 20 (see arrows). Randomly spaced \( T \) (dashed light line) yields slightly higher rates for smaller values of \( T \).

Figure 3.22: Sensitivity rates for freq = 3.8, \( T = 50 \), as a function of \( \eta \) in finer detail. Evenly (solid dark line) and randomly (dashed light line) spaced large \( T \) produce similarly decreasing rates.
3.2.3 Performance using regression-based $f$

The results from the regression based simulation were similar to the Haar and sinusoidal simulations in that as $p$ increased from 0.1 to 0.9, there were only slight improvements in the sensitivity and observed false discovery rates. Therefore, we will again present only those results from the cases where $p = 0.1$. The functions upon which our algorithm performed the best were $f_{reg1}(t)$ and $f_{reg4}(t)$. In general, sensitivity and false discovery rates were better when $T$ was evenly spaced. However, when $T$ was small, we again saw that randomly spaced $T$ yielded some improvements.

For regression function $f_{reg1}(t) = 1 - 48t + 218t^2 - 315t^3 + 145t^4$ (see Tab. 3.9 and Fig. 3.23), a minimal level of added noise ($\eta = 0.1$) required 30 evenly spaced time points to yield a 74% sensitivity rate ($FDR_{obs} = 0.10$), and 40 randomly spaced time points for 72% sensitivity ($FDR_{obs} = 0.08$). If evenly spaced $T$ was reduced to 25, sensitivity decreased to 30% ($FDR_{obs} = 0.19$). And if randomly spaced $T$ was reduced to 30, sensitivity decreased to 18% ($FDR_{obs} = 0.22$).

For large $T$, a small amount of added noise adversely affected sensitivity rates. For evenly spaced $T = 50$, sensitivity rates dropped from 99% ($FDR_{obs} = 0.07$) to 49% ($FDR_{obs} = 0.09$) as $\eta$ increased from 0.25 to 0.50. For randomly spaced $T = 50$, sensitivity dropped from 90% ($FDR_{obs} = 0.12$) to 51% ($FDR_{obs} = 0.09$) as $\eta$ increased from only 0.25 to 0.50.
Table 3.9: Observed sensitivity rates for $f_{regr1}(t)$, evenly spaced $T$ and $p = 0.1$

<table>
<thead>
<tr>
<th>$T$</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\eta$</td>
<td>2.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.03</td>
<td>0.02</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.08</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0.05</td>
<td>0.01</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.05</td>
<td>0.06</td>
<td>0.10</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.13</td>
<td>0.21</td>
<td>0.30</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.19</td>
<td>0.30</td>
<td>0.74</td>
<td>0.88</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.23: Surface plots (b and c) of observed sensitivity rates from simulations generated from regression function (a) $f_{regr1}(t) = 1 - 48t + 218t^2 - 315t^3 + 145t^4$.

To get a better idea of what was going on in between our standard simulation values of $T$ and $\eta$, we ran additional simulations using small $\eta$ with $T$ in finer detail, and large $T$ with $\eta$ in finer detail. The results of these simulations are presented in Figures 3.24 and 3.25. We can see from Fig. 3.24, which shows how $T$ (with minimal added noise) affects sensitivity, that the largest jump occurs for both evenly and randomly spaced time points when $T$ is between 26 and 28, and that evenly spaced $T$ outperforms randomly spaced $T$ near 30 time points.

Fig. 3.25 shows how $\eta$ (with large $T$) affects sensitivity; we see that the biggest difference between evenly and randomly spaced $T$ occurs between $\eta = 0.4$
and \( \eta = 0.6 \), which is also where the sensitivity rates drop the most. Otherwise, they decrease similarly.

![Graph showing sensitivity rates for \( f_{reg1}(t) \), \( \eta = 0.1 \), as a function of \( T \) in finer detail. Evenly spaced \( T \) (solid dark line) differs most from randomly spaced \( T \) (dashed light line) near \( T = 30 \). Rates decrease significantly for both once \( T \) reaches the mid-twenties.](image)

**Figure 3.24:** Sensitivity rates for \( f_{reg1}(t) \), \( \eta = 0.1 \), as a function of \( T \) in finer detail. Evenly spaced \( T \) (solid dark line) differs most from randomly spaced \( T \) (dashed light line) near \( T = 30 \). Rates decrease significantly for both once \( T \) reaches the mid-twenties.

![Graph showing sensitivity rates for \( f_{reg1}(t) \), \( T = 50 \), as a function of \( \eta \) in finer detail. Evenly spaced \( T \) (solid dark line) and randomly spaced \( T \) (dashed light line) produce similarly decreasing rates except for \( 0.4 \leq \eta \leq 0.6 \), where randomly spaced \( T \) yields smaller sensitivity rates.](image)

**Figure 3.25:** Sensitivity rates for \( f_{reg1}(t) \), \( T = 50 \), as a function of \( \eta \) in finer detail. Evenly spaced \( T \) (solid dark line) and randomly spaced \( T \) (dashed light line) produce similarly decreasing rates except for \( 0.4 \leq \eta \leq 0.6 \), where randomly spaced \( T \) yields smaller sensitivity rates.
Table 3.10: Observed sensitivity rates for $f_{regr2}(t)$, randomly spaced $T$ and $p = 0.1$

<table>
<thead>
<tr>
<th>$T$</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td>0.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
<td>0.01</td>
<td>0.02</td>
<td>0.06</td>
</tr>
<tr>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.05</td>
<td>0.11</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0.08</td>
<td>0.15</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Figure 3.26: Surface plots (b and c) of observed sensitivity rates from simulations generated from regression function (a) $f_{regr2}(t) = 0.3e^{-64(t-0.25)^2} + 0.7e^{-256(t-0.75)^2}$.

Simulations using regression function $f_{regr2}(t) = 0.3e^{-64(t-0.25)^2} + 0.7e^{-256(t-0.75)^2}$ (see Tab. 3.10 and Fig. 3.26), yielded the poorest results out of all four regression functions. Results were most similar to those from simulations using function $f = sin(6.0 \pi t)$. Evenly spaced $T$ yielded sensitivity rates near zero even for large $T$ and small $\eta$. The results from randomly spaced $T$ are shown in Tab. 3.10. For the best case scenario ($T = 50$ and $\eta = 0.1$), randomly spaced $T$ produced a sensitivity rate of 47% ($FDR_{obs} = 0.11$), while evenly spaced $T$
produced a sensitivity rate of only 11% ($FDR_{obs} = 0.21$).

Regression function $f_{regr3}(t) = 10e^{-10t}$, (see Tab. 3.11 and Fig. 3.27), yielded better results than $f_{regr2}(t)$ but were still not as good as those produced using $f_{regr1}(t)$.

**Table 3.11:** Observed sensitivity rates for $f_{regr3}(t)$, evenly spaced $T$ and $p = 0.1$

<table>
<thead>
<tr>
<th>$T$</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0200</td>
</tr>
<tr>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0100</td>
<td>0.0300</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0100</td>
<td>0</td>
<td>0.0100</td>
<td>0</td>
<td>0.0200</td>
<td>0</td>
<td>0</td>
<td>0.0200</td>
</tr>
<tr>
<td>0.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0200</td>
<td>0</td>
<td>0.0200</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0100</td>
<td>0.0200</td>
<td>0.0500</td>
<td>0.0600</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0100</td>
<td>0.0400</td>
<td>0.1500</td>
<td>0.4800</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0600</td>
<td>0.0200</td>
<td>0.5000</td>
<td>0.9000</td>
</tr>
</tbody>
</table>

![Figure 3.27](image)

**Figure 3.27:** Surface plots (b and c) of observed sensitivity rates from simulations generated from regression function (a) $f_{regr3}(t) = 10e^{-10t}$.

The smallest amount of added noise ($\eta = 0.1$) required 40 evenly spaced time points to yield a 59% sensitivity rate ($FDR_{obs} = 0.12$), and 40 randomly spaced time points for 63% sensitivity ($FDR_{obs} = 0.12$). If evenly spaced $T = 30$, sensitivity decreased to 2% ($FDR_{obs} = 0.50$). And if randomly spaced $T = 30$, sensitivity...
sensitivity decreased to 11% ($FDR_{obs} = 0.21$).

Even for large $T$, it didn’t take much added noise to adversely affect sensitivity rates. For evenly spaced $T = 50$, sensitivity rates dropped from 99% ($FDR_{obs} = 0.08$) to 48% ($FDR_{obs} = 0.06$) as $\eta$ increased from 0.10 to 0.25. For randomly spaced $T = 50$, sensitivity dropped from 75% ($FDR_{obs} = 0.06$) to 35% ($FDR_{obs} = 0.12$) as $\eta$ increased from only 0.10 to 0.25.

Simulations using regression function $f_{regr4}(t)$ (see Tab. 3.12 and Fig. 3.28) yielded the best observed sensitivity and false discovery rates of all four re-

<table>
<thead>
<tr>
<th>$T$</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\eta$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
</tr>
<tr>
<td>1.5</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
</tr>
<tr>
<td>0.75</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
</tr>
<tr>
<td>0.25</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
</tr>
<tr>
<td>0.1</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0100</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
<td>0.0300</td>
</tr>
</tbody>
</table>

Figure 3.28: Surface plots (b and c) of observed sensitivity rates from simulations generated from regression function (a) $f_{regr4}(t) = e^{k(t-\frac{1}{3})}$, where $k = -2$ for $t < \frac{1}{3}$ and $k = 1$ for $t \geq \frac{1}{3}$. 

---

57
gression based generating functions. Using the smallest level of added noise ($\eta = 0.1$), 20 evenly spaced time points produced a 100% sensitivity rate ($FDR_{obs} = 0.07$), and 20 randomly spaced time points produced 79% sensitivity ($FDR_{obs} = 0.08$). If evenly spaced $T = 15$ however, sensitivity decreased to 38% ($FDR_{obs} = 0.24$). And if randomly spaced $T = 15$, sensitivity decreased to 32% ($FDR_{obs} = 0.06$).

For large $T$, added noise did not affect sensitivity rates until $\eta = 1.0$. For evenly spaced $T = 50$, sensitivity rates dropped from 81% ($FDR_{obs} = 0.08$) to 40% ($FDR_{obs} = 0.17$) as $\eta$ increased from 0.75 to 1.0. For randomly spaced $T = 50$, sensitivity dropped from 67% ($FDR_{obs} = 0.14$) to 28% ($FDR_{obs} = 0.03$) as $\eta$ increased from 0.75 to 1.0.

We ran additional simulations using small $\eta$ with $T$ in finer detail, and large $T$ with $\eta$ in finer detail, the results of which are presented in Figures 3.29 and 3.30. We can see from Fig. 3.29, which shows how $T$ (with minimal added noise) affects sensitivity, that evenly and randomly spaced $T$ differ most between 22 and 16. Additionally, while rates for randomly spaced $T$ decrease gradually for $T < 15$, rates for evenly spaced $T$ appear to favor an even number of time points, falling when there are an odd number.

Fig. 3.30 shows how $\eta$ (with large $T$) affects sensitivity; we see that the biggest difference between evenly and randomly spaced $T$ occurs between $\eta = 0.7$ and $\eta = 1.0$, which is also where the sensitivity rates drop the most. Otherwise, they decrease similarly.
Figure 3.29: Sensitivity rates for $f_{reg4}(t)$, $\eta = 0.1$, as a function of $T$ in finer detail. Rates for randomly spaced $T$ (dashed light line) decrease more gradually than evenly spaced $T$ when the number of time points is less than 15. And is evenly spaced $T$ favoring an even number of time points when $T$ is small?

Figure 3.30: Sensitivity rates for $f_{reg4}(t)$, $T = 50$, as a function of $\eta$ in finer detail. Evenly (solid dark line) and randomly (dashed light line) spaced large $T$ produce similarly decreasing rates except for $0.7 \leq \eta \leq 1.0$, where randomly spaced $T$ yields smaller sensitivity rates.
3.3 Results: Microarray Gene Expression Time Series

We applied our algorithm to the microarray gene expression time series data obtained from the mammary tissue of rattus norvegicus over the four day estrous cycle. We experimented with a number of different approaches. Each approach was performed 10 times, so that we might see whether a gene was selected every time, none of the time, or some of the time. Since the total number of time points happened to already be a power of 2, we first used a basis size of 32 ($2^5$). We ran the data through the algorithm using basis sizes of 16, 8, and 4 as well. This approach, with basis sizes of 8, 16, and 32 yielded the most promising results; several genes were found to be significant for each of the ten rounds.

The other approaches involved using only subsets of the series in an attempt to fashion a more uniform sampling rate. The vector of time points was uniform within each of the four days, between 7AM and 7PM. But there were three large gaps of time, between 7PM and 7AM, where no samples were taken (see Fig. 3.31). Keeping only the first and last time point of each day yielded 8 rather

![Figure 3.31](image)

Figure 3.31: A plot indicating the time points sampled. Each asterisk marks the time (in hours) that a tissue sample was taken from a rat specimen across the 4 day estrous cycle.
uniform sample points across the entire time window. However, applying our algorithm to this subset of the dataset (using a basis size of 8 or 16) yielded poor results; only a few genes were tagged, and none of them were selected more than 50% of the time.

Additionally, we applied the algorithm to each of the four days (estrus cycle phases) separately, using a basis size of 8 and then 16. These four subsets also consisted of 7 to 9 uniformly spaced time points, and also yielded poor results. As we saw in the simulation testing, for evenly spaced small $T$, results tended to not be good, even with a minimal amount of added noise. The poor results here support the indication that 8 sample points, although highly uniform, are not enough to enable the detection of temporal patterns. We will limit our reporting therefore, to the results obtained when using the entire dataset.

Among the 21,440 genes examined, 40 genes were consistently tagged as exhibiting significant temporal patterns across the four days of the estrous cycle. Fifty genes were found to be significant 90% of the time, and another 53 were tagged 80% of the time. Tab. 3.13 gives the frequency distribution of these significance rates. Descriptions of the "top forty" are featured in Tab. 3.14.

<table>
<thead>
<tr>
<th>Rate of Significance</th>
<th>1.0</th>
<th>0.9</th>
<th>0.8</th>
<th>0.7</th>
<th>0.6</th>
<th>0.5</th>
<th>0.4</th>
<th>0.3</th>
<th>0.2</th>
<th>0.1</th>
<th>0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Count</td>
<td>40</td>
<td>50</td>
<td>53</td>
<td>53</td>
<td>70</td>
<td>90</td>
<td>118</td>
<td>157</td>
<td>219</td>
<td>492</td>
<td>19702</td>
</tr>
</tbody>
</table>

A search of the literature revealed additional information on some of the forty genes consistently identified by our algorithm.
The aortic preferentially expressed gene 1 (APeg1; Ref# 1 in Tab. 3.14), which is expressed in arterial smooth muscle cells, was found by our algorithm to exhibit a temporal pattern across the estrous cycle. Research has shown that cycles of aortic oxygen absorption are correlated with phases of the estrous cycle in female rats [23].

![Graph showing normalized expression signal for APeg1 over the estrous cycle.](image)

**Figure 3.32:** The Haar wavelet reconstruction of the normalized expression signal for APeg1. An increase in expression appears in the estrus and metestrus phases.

The Haar wavelet reconstruction of the expression signal (a linear combination of the wavelet coefficients and basis functions) is shown in Fig. 3.32. Expression appears to increase in the estrus and metestrus phases. Honda et al. found that relaxation of the aorta was greater for rats in the estrus phase, than for those in the proestrus, metestrus or diestrus phases [15]. Relaxation of the aorta allows blood to flow more smoothly throughout the body thereby facilitating oxygen absorption. No published works were found involving APeg1.
expression in mammary tissue.

The gene associated with mammary cancer in rats (RMT-1; Ref. #7 in Tab. 3.14), was found by our algorithm to be temporally significant within the estrous cycle. Researchers at UC Berkeley were the first to identify this gene and its association with rat mammary tumors [10]. Their research has shown RMT-1 to be highly expressed in the normal breast cells of virgin rats, and even more so in rat mammary cancers. No previously published works were found linking this gene to the estrous cycle however. The Haar wavelet reconstruction of this gene's expression signal is shown in Fig. 3.33. Expression levels are high in the first two phases, and are low in the last two.

![Graph showing expression levels over estrous cycle phases](image)

**Figure 3.33:** The Haar wavelet reconstruction of the normalized expression signal for RMT-1. Expression values are high in the first half of the estrous cycle, and low in the second half.

Embryo receptivity, embryotrophy as well as the immune system are all partially regulated by cytokines, which are a group of proteins and peptides.
used in cell signaling. Interleukins make up a subset of cytokines. The mRNA expression of inflammatory cytokines interleukin-4 (*IL-4*) and interleukin-6 (*IL-6*) was found to be expressed in the corpus luteum of the porcine (pig) [29]. The corpus luteum is a temporary progesterone-producing structure present in the ovary only during the metestrus and diestrus phases of the estrous cycle. We would expect that genes highly expressed in the corpus luteum would therefore show higher levels of activity during these phases.

![Graph](image)

**Figure 3.34:** The Haar wavelet reconstruction of the normalized expression signal for IL-17B. Expression is low in the first half of the cycle, and high during the last two phases.

The cytokine interleukin-17B (*IL-17B*; Ref# 18 in Tab. 3.14) was found to have significant temporal expression patterns by our algorithm. IL-17B is a member of the subfamily of interleukins, IL-17, which is most notably involved in governing inflammatory responses and the production of many other cytokines (such as IL-6). The Haar wavelet reconstruction of the expression activity of IL-
17B is shown in Fig. 3.34. Expression levels are low in the first two phases, and are high during the metestrus and diestrus phases, coinciding with the presence of the corpus luteum.

The angiogenesis inhibiting Thrombospondin proteins (TSP) are expressed in the normal and hyperplastic human breast, and have been of interest in breast cancer research due to their anti-tumorigenic functions. Thrombospondins are a family of multi-functional proteins and can be divided into two subfamilies: A, containing TSP-1 and -2, and B, containing TSP-3, -4 and -5. Our algorithm consistently identified as temporally significant the gene responsible for Thrombospondin-2 \((TSP-2; \text{Ref# 26 in Tab. 3.14})\). In reference to estrous,

![Graph](image-url)

**Figure 3.35:** The Haar wavelet reconstruction of the expression signal for TSP2. Expression is higher in the first half, and lower in the second half of the estrous cycle.

TSP was found to be expressed in the preluteal follicles (present in the proestrus phase) and the mature corpus luteum (present in the metestrus phase) of the
rat [4]. The Haar wavelet reconstruction of the expression activity of TSP-2 is shown in Fig. 3.35. Expression levels are higher in the proestrus and estrus phases, and are lower during the second half of the cycle.

Laminin (LN) is a component of thin connective membranes (such as ovarian follicles) and is involved in such functions as cell growth, differentiation and migration. Research has suggested that LN plays a significant role in follicular development [21, 2], which occurs during the proestrus and estrus phases of the estrous cycle. The reconstructed expression signal of Laminin α-5 (LAMA5; Ref# 30 in Tab. 3.14) is shown in Fig. 3.36. Except for a few neighboring samples in the estrus phase resulting in high levels of mRNA concentration, the expression levels for LAMA5 appear to moderately decrease from the beginning to the end of the cycle.

**Figure 3.36:** The Haar wavelet reconstruction of the expression signal for LAMA5. Expression moderately decreases from the beginning to the end of the cycle, with some high level readings occurring in the estrus phase.
In reference to its link to mammary tissue, Laminin α-5 (LAMA5; Ref# 30 in Tab. 3.14) was found to be differentially expressed in metastatic tumor cells [32].

Table 3.14: Genes found to to have temporal significance 100% of the time

<table>
<thead>
<tr>
<th>Ref#</th>
<th>Gene ID</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1368001_at</td>
<td>APEG1</td>
<td>Aortic preferentially expressed gene 1</td>
</tr>
<tr>
<td>2</td>
<td>1368470_at</td>
<td>Gamma-glutamyl hydrolase</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1368567_at</td>
<td>MIPPE5 protein</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1369132_at</td>
<td>SLC18A2</td>
<td>Solute carrier family 18, member 2</td>
</tr>
<tr>
<td>5</td>
<td>1369669_at</td>
<td>Neurolysin (metallopeptidase M3 family)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1370317_at</td>
<td>KCR1</td>
<td>Potassium channel regulator 1</td>
</tr>
<tr>
<td>7</td>
<td>1370646_at</td>
<td>RMT-1</td>
<td>Mammary cancer associated protein RMT-1</td>
</tr>
<tr>
<td>8</td>
<td>1370805_at</td>
<td></td>
<td>Melanocyte-specific gene 1 protein</td>
</tr>
<tr>
<td>9</td>
<td>1371830_at</td>
<td></td>
<td>Similar to ubiquitin-like 1 (sentrin) activating enzyme subunit 1; ubiquitin-like (sentrin) activating enzyme E1A; SUMO-1 activating enzyme subunit 1; DNA segment; Chr 7, ERATO Doi 177, expressed</td>
</tr>
<tr>
<td>10</td>
<td>1372120_at</td>
<td></td>
<td>Transcribed locus</td>
</tr>
<tr>
<td>11</td>
<td>1372281_at</td>
<td></td>
<td>Similar to hypothetical protein MGC28394</td>
</tr>
<tr>
<td>12</td>
<td>1373656_at</td>
<td></td>
<td>Similar to RIKEN cDNA 1110014L17</td>
</tr>
<tr>
<td>13</td>
<td>1374260_at</td>
<td></td>
<td>UL-R-CM0-bji-d-03-0-UI-a! Rattus norvegicus cDNA clone UL-R-CM0-bji-d-03-0-UI 3', mRNA sequence.</td>
</tr>
<tr>
<td>14</td>
<td>1374594_at</td>
<td></td>
<td>Similar to RIKEN cDNA 1600022D21</td>
</tr>
<tr>
<td>15</td>
<td>1375091_at</td>
<td></td>
<td>Similar to KIAA0192 protein</td>
</tr>
<tr>
<td>16</td>
<td>1375469_at</td>
<td></td>
<td>SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4</td>
</tr>
<tr>
<td>17</td>
<td>1375637_at</td>
<td></td>
<td>Similar to RIKEN cDNA 1110003E01</td>
</tr>
<tr>
<td>18</td>
<td>1376445_at</td>
<td>IL-17B</td>
<td>Interleukin 17B</td>
</tr>
<tr>
<td>19</td>
<td>1377469_at</td>
<td></td>
<td>Transcribed locus</td>
</tr>
<tr>
<td>20</td>
<td>1377982_at</td>
<td></td>
<td>Similar to hypothetical protein D90042H13</td>
</tr>
<tr>
<td>21</td>
<td>1381130_at</td>
<td></td>
<td>UL-R-BT1-bmy-a-11-0-UI-a! Rattus norvegicus cDNA clone UL-R-BT1-bmy-a-11-0-UI 3', mRNA sequence</td>
</tr>
<tr>
<td>22</td>
<td>1382187_at</td>
<td></td>
<td>Similar to RIKEN cDNA 286022K21</td>
</tr>
<tr>
<td>23</td>
<td>1383930_at</td>
<td></td>
<td>Similar to Ribonuclease P protein subunit p38 (RNaseP protein p38)</td>
</tr>
<tr>
<td>24</td>
<td>1383424_at</td>
<td></td>
<td>Similar to thymidylate kinase family LPS-inducible member; thymidylate kinase homologue</td>
</tr>
<tr>
<td>25</td>
<td>1384836_at</td>
<td></td>
<td>UL-R-Co-jp-b-06-0-UI-a! Rattus norvegicus cDNA clone UL-R-Co-jp-b-06-0-UI 3', mRNA sequence</td>
</tr>
</tbody>
</table>
Table 3.13: (Cont.)

<table>
<thead>
<tr>
<th>Ref#</th>
<th>Gene ID</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>1388751.at</td>
<td>TSP-2</td>
<td>Thrombospondin 2</td>
</tr>
<tr>
<td>27</td>
<td>1385784.at</td>
<td>Similar to calcium-activated chloride channel CLCA4</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>1387664.at</td>
<td>ATPase, H+ transporting, lysosomal (vacuolar proton pump), beta 56/58 kDa, isoform 2</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>1387868.at</td>
<td>SLC7A7</td>
<td>Solute carrier family 7 (cationic amino acid transporter, y+system), member 7</td>
</tr>
<tr>
<td>30</td>
<td>1388932.at</td>
<td>LAMA5</td>
<td>Laminin, alpha 5</td>
</tr>
<tr>
<td>31</td>
<td>1389173.at</td>
<td>Similar to SON protein</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>1389411.at</td>
<td>Transcribed locus</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>1389439.at</td>
<td>Transcribed locus</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>1389612.at</td>
<td>Transcribed locus, highly similar to XP.2432369.1 similar to RIKEN cDNA 2400002D02 (Rattus norvegicus)</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>1390152.at</td>
<td>Transcribed locus</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>1391382.at</td>
<td>Transcribed locus</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>1394577.at</td>
<td>Transcribed locus</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>1395387.at</td>
<td>Similar to RIKEN cDNA 1600029D21</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>1397729.x.at</td>
<td>Similar to RIKEN cDNA 1600029D21</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>1398661.at</td>
<td>UI-R-CVO-brr-e-10.0-UI.x</td>
<td>UI-R-CVO Rattus norvegicus cDNA clone UI-R-CVO-brr-e-10.0-UI 3', mRNA sequence.</td>
</tr>
</tbody>
</table>
4. SUMMARY AND CONCLUSIONS

There were several questions we hoped to be able to answer based on the results of our simulations. In this chapter, we will summarize the results from the previous chapter and address the questions we posed about how the size, rate, and response variation of a sample affects the results of our methodology.

4.1 Sample Size

One of the primary questions we hoped to answer from our simulations was how many sample points are needed to determine the presence of a temporal pattern. In general, the less noise present in the sample response, the smaller the sample size could be - to a point. Usually, no matter how small the amount of noise, any series with less than 8 sample points did not yield any discoveries, true or not.

When noise was only a small factor, we found that the answer to this question greatly depended upon the characteristics of the underlying signal. Our simulation results indicated that lower frequencies fared better, despite sample sizes as small as 20. Underlying signals such as these — including the fourth regression function, sinusoidal functions with frequencies near 0 or 2, and the Haar wavelet step function — consistently yielded sensitivity rates better than 90% and observed false discover rates smaller than 0.1.

The underlying function whose sample size was least susceptible to a large degree of noise was the Haar step function; when the standard deviation of
added Gaussian noise equalled the maximum amplitude of the signal ($\eta = 1.0$), the sensitivity rate for evenly spaced $T = 25$ was 76%, for randomly spaced $T$ 60%, both with observed false discovery rates less than or equal to 0.1.

4.2 Variation of Sample Response

Another primary question we wanted to address was what degree of noise, or sample response variation, could be tolerated in identifying the presence of a temporal pattern. In general, the more sample points, the more that noise could be tolerated - but also only to a point. Usually, even for a sample size as large as 50, any series with additional noise having as much as double the maximum amplitude of the underlying signal yielded sensitivity rates of less than 5%.

When sample size was large, we found that the answer to this question appeared to be mostly independent of the underlying signal. Our simulation results indicated that for nearly all underlying functions simulated, it was when $\eta$ was between 0.5 and 1.0 that sensitivity rates began to decrease severely.

The underlying function which best tolerated a high level of variation in the sample response was again the Haar step function. When the standard deviation of added Gaussian noise doubled the maximum amplitude of the signal ($\eta = 2.0$), the sensitivity rate for evenly spaced $T = 50$ was 46%, for randomly spaced $T$ 31%, both with observed false discovery rates less than 0.1.

4.3 Sampling Rate

Another question we hoped to answer was how irregular sampling affected the methodology. For larger $T$ (usually $T > 15$), evenly spaced time points produced moderately better results than randomly spaced samples. When $T$
was small, randomly spaced $T$ sometimes yielded better results than evenly spaced.

Notable was that for a large number of sample points, results did not differ greatly between randomly and evenly spaced time points as the level of added noise changed. This can be seen in the Results chapter figures (for example, see Figures 3.22 and 3.30) which plotted sensitivity rates as a function of $\eta$.

4.4 Frequency of the Underlying Temporal Response

Since wavelets do not detect stationary frequencies, it was not surprising that the higher the sine frequency, the poorer the results. But why then were results from $\text{frq} = 3.8$ better than those from $\text{frq} = 2.5$? While results for $\text{frq} = 5.2$ were better than those for $\text{frq} = 6$ (because of the lower frequency), results for $\text{frq} = 8$ were actually better than when $\text{frq} = 5.2$, despite the lower frequency.

These jumps of improvement when frequency was at or near a power of two was a result of the underlying function mimicking the scales of the step function. Fig. 2.1 shows the Haar wavelet scales $k = 2, \ldots, 8$. Just as the algorithm performed well on the basic Haar step function (Fig. 2.1(a)), so it did for the curvier version, the sine wave with frequency 2. Similarly, the sine wave with frequency 4 is aligned with the scaled Haar step functions in Fig. 2.1(b and c), and the sine wave with frequency 8 is aligned with the scaled Haar step functions in Fig. 2.1(d-g).

The pattern of decreasing sensitivity rates as frequency increased also held true within the power-of-two frequencies (e.g., $\text{frq} = 8$ was not as successful as $\text{frq} = 4$, nor $\text{frq} = 4$ as successful as $\text{frq} = 2$).
4.5 Which Wavelet Coefficient Performs Best?

Of the 7 wavelet coefficients assessed ($C_2, \ldots, C_6$), did any stand out as being more or less often the coefficient responsible for producing a significantly small $p$-value? Recall that it is the smallest of the seven coefficient-generated $p$-values ultimately associated with each gene. And then the FDR procedure determines which of these minimum $p$-values is significantly small. We ran the Haar Step function-based simulation in which we recorded which $C_k$ was responsible for a positive. Tab. 4.1 shows the results (averaged over 10 runs) for each $C_k$ from optimal parameters $T = 50$ and $\eta = 0.1$. Clearly, for very low frequencies like

<table>
<thead>
<tr>
<th>Table 4.1: Percentage Distribution of Positive-Responsible $C_k$ (when $T = 50$ and $\eta = 0.1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>False Positives (T uniform)</td>
</tr>
<tr>
<td>False Positives (T random)</td>
</tr>
<tr>
<td>True Positives (T uniform)</td>
</tr>
<tr>
<td>True Positives (T random)</td>
</tr>
</tbody>
</table>

the Haar Step function, if there are a relatively large number of time points with a relatively small degree of noise, $C_2$ is consistently the responsible coefficient for correct discoveries.
5. FUTURE WORK

The incorporation of positive-responsible $C_k$ considerations and whether they are of use in detecting temporal significance in microarray time series could be a fruitful avenue of exploration. For example, is $C_2$ usually positive-responsible even for high frequency signals? What about when there are a smaller number of time points? Is it possible that there are scenarios in which a certain coefficient is often incorrectly positive-responsible? A more complete understanding of the sensitivities of particular coefficients may be useful in the detection of time series expression signals having specific characteristics, such as low, or high frequency.

Our use of the Haar wavelet was a simplified approach. Future research might incorporate the use of continuous wavelets, such as the Morlet, Meyer, or Mexican hat wavelets, in the analysis of microarray gene expression time series.

Others who might implement a wavelet transform in the analysis of microarray time series, may find it useful to investigate other methods of standardization of the data. Our approach of normalization involved taking the log and centering each signal about zero. Dividing by the 1-norm of the signal vector may have provided better results for certain types of signals, such as those simulated using the second regression function. Making the length of the data vectors equal allows small and large scale activity to exist within the same scale.

In order to best test such an algorithm as presented in this paper, it is important to be able to replicate in simulations the nuances of actual microarray
time series. It would be informative therefore to experiment with different types of added noise. For example, incorporate noise that is linked to groups of signals and varies with the time points. This could serve to simulate clusters of genes which behave similarly (e.g., always upregulated, downregulated, or neither) from one organism (time point) to another.

It could be highly informative to apply this approach to microarray gene expression time series in which temporally significant genes are already known. An integrated analysis of the extensive and well-studied microarray time series (by Cho and Spellman, 1998) on gene expression in yeast (Saccharomyces cerevisiae) could help to confirm the success of any novel approach.

And finally, further investigation into genes found temporally significant by our algorithm could reveal underlying genetic functions within the estrous cycle.

Along with many other works, our results support the importance of maximizing the number of time points sampled when performing a microarray time series experiment. Even when faced with a relatively large amount of sample variation response, a larger number of time points, whether uniformly taken or not, was usually able to preserve an existing temporal pattern.
REFERENCES


