A SYSTEMS BIOLOGY APPROACH FOR ASSESSING CORTICOSTEROID ACTIVITY

By

ERIC H. YANG

A dissertation submitted to the Graduate School – New Brunswick

Rutgers, The State University of New Jersey

In Partial Fulfillment of the requirements For the Degree of

Doctor of Philosophy

Graduate Program in Biomedical Engineering

Written under the direction of Dr. Ioannis P. Androulakis

And Approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2008

ABSTRACT OF THE DISSERTATION

A Systems Biology Approach for Assessing Corticosteroid Activity

By Eric H. Yang

Dissertation Director: Dr Ioannis P. Androulakis

Glucocorticosteroids are endogenous hormones that are produced in the adrenal gland and are responsible for the regulation of glucose metabolism. However, corticosteroids are also utilized therapeutically for the suppression of the immune system as well as the treatment of inflammation. Despite their immunosuppressive and anti-inflammatory role, long term use of corticosteroids is problematic due to severe side effects. Deciphering a comprehensive mechanism of corticosteroid activity might offer valuable clues for mitigating the adverse side effects.

To accomplish this task we have leveraged high throughput experimental methods such as mRNA microarrays as well as a new technique for measuring transcription factor activity based on a novel technology, namely the Living Cell Array. However, due to the large amount of information which is generated by these methods, it is imperative that automated methods be developed to identify the critical pieces of information present in the large amount of data generated. Our primary focus was on characterizing the dynamics of transcriptional response and to further isolate relevant regulatory structures.

Our analysis is based on in vivo (liver) measurements of a rat model as well as hepatocyte cultures in the living cell array. Major findings of this work suggest that the anti-inflammatory effect of corticosteroids appear to be a transient phenomenon under a sustained infusion of corticosteroids, whereas the metabolic effects of corticosteroids appear to be ongoing in concordance with sustained stimuli via the drug. Furthermore, it is suggested that the monomeric form or the glucocorticosteroid receptor was active. Specifically, given the consensus sequence associated with the dimeric form of the corticosteroid receptor, there is evidence that a monomeric form of the glucocorticosteroid will bind, and more importantly cause the transcription of said gene. Finally, evidence is obtained that the anti-inflammatory effects of corticosteroids aside from being transient are regulated by a feedback loop.

A model of corticosteroid activity is finally proposed which utilizes two separate and active forms of the glucocorticosteroid receptor. The advantage of this model over the currently accepted one is that it is able to replicate the response of the system to corticosteroids with a very simple mechanistic explanation. The model recreates both the response of the host to an injection as well as infusion of corticosteroid, as well as the circadian variation of gene expression due to the circadian oscillations of endogenous corticosteroid levels.

Acknowledgement

I'd like to thank whoever was on the Rutgers Biomedical Engineering Admissions Committee for accepting me when things were looking up for me. And I'd like to thank my parents John and Fenli Yang and my brother Roger for convincing me to go back to school.

Table of Contents

List of Tables

List of Illustrations

Introduction

Corticosteroids are a class of steroid derived hormones used to control processes such as glucose utilization and ion balance[1]. They are normally separated into two classes, mineralcorticosteroids and glucocorticosteroids with mineralcorticosteroids responsible for mediating ion balance and glucocorticosteroids responsible for mediating glucose metabolism. In this dissertation, we will use the term corticosteroids to refer solely to glucocorticosteroids.

While corticosteroids were discovered to regulate the levels of glucose utilization, synthetic corticosteroids such as methylprednisolone are used therapeutically for the treatment of chronic inflammation as well as the suppression of the immune system[2]. For these two roles, corticosteroids are very effective. However, their therapeutic index is quite low due to the presence of a large number of serious side effects such as muscle wasting, metabolic shift, and steroid induced diabetes[3, 4]. Because of these severe side effects, the long term use of corticosteroids must be carefully weighed against the negative side effects associated with them[5]. However, if one were to be able to reduce the side effects associated with corticosteroids, it would open up a new avenue for the prolonged treatment of inflammation and allow for more effective therapies such as suppressing the immune system after organ transplantation. However, before it is possible to establish possible methods for mitigating the side effects of corticosteroids, one must first determine the underlying mechanism by which corticosteroids exert their influence upon biological system. The importance of deriving a mechanism is that a mechanism represents a set of hypotheses from which rational strategies for intervention can be obtained.

The fact that prolonged exposure to corticosteroids results in significant and severe side effects is not all that surprising given that their primary biological role is the mediation of an important metabolic process[6]. However, given their effectiveness at controlling inflammation and suppressing the immune system, the question is whether inflammation and the immune response is intrinsically linked to metabolism, and as such whether it is possible to suppress the immune response and inflammation without having some impact upon metabolism. The ability of non-steroid anti-inflammatory drugs (NSAIDS) to target inflammation without the severe side effects of corticosteroids suggests that this may be possible[7]. However, the lack of severe metabolic side effects of NSAIDS may be a dose dependent phenomenon because of the increased effectiveness of corticosteroids at treating inflammation as compared to NSAIDS at lower doses. Therefore, it may be possible that if a dose of NSAIDS were high enough to have the same anti-inflammatory effect as corticosteroids that similar metabolic effects may be seen.

Corticosteroids are thought to function via transcriptional regulation, with the immuno-suppressive effects arising due to the inhibition of cytokines such as IL-1, IL-6, and IFN-γ[8] and the anti-inflammatory effect arises from the regulation of lipocortin which decreases the levels of the inflammatory agent phospholipase A2[9]. The common thread between these two mechanisms is the fact that corticosteroids act transcriptionally, and not directly. The current fifth generation model of corticosteroids which has been proposed by our collaborators Almon et al.[10], hypothesize that corticosteroids will bind to a Glucocorticosteroid Receptor (GR), which then dimerizes, and translocates into the nucleus. After the translocation event, the activated receptor binds to a GRE element up-stream of a given gene and either up or down-regulates the expression of that gene.

This drug effect was described as an indirect effect model[11], in which the primary effect is mediated through the production of secondary proteins. The critical aspect of this model is the fact that the drug activity no longer directly correlates with the amount of the drug that is currently within the system. Therefore, the drug could be having its maximal effect long after the drug has been cleared from the circulation. Therefore, the pharmacokinetics of a given drug no longer has a one to one correlation with the pharmacodynamics of the drug. Furthermore, the responses of a given drug are more complex because by triggering a transcriptional cascade, it is not clear as to when the effect of the drug end, and where the compensatory mechanisms associated with the response are beginning.

Because the drug acts indirectly and regulates the production of proteins, rather than measuring how the levels of various metabolites change, we can instead focus primarily upon changes in mRNA gene expression levels. The levels of mRNA will be used as a surrogate for the levels of proteins which are the final effectors of drug activity. The fact that the mRNA expression levels can be measured in parallel easily through the use of mRNA microarrays allows us to utilize a systems biology approach by allowing us to first identify a set of relevant systems, draw connections between a large number of systems, and finally test hypotheses *in silico* without requiring one to necessarily run different experiments.

Current work into the activity of corticosteroid activity has yielded some tentative ideas as to what systems are affected along with a preliminary mechanism for corticosteroid activity. In the current state of the art, genes which were selected as up or down-regulated had their functionality identified. Secondly, the expression profiles for selected marker genes were used to create PK/PD models that described corticosteroid activity. The fifth generation model of corticosteroid activity consists of 6 sets of differential equations which describe the activity of six

groups of genes **Figure 1**. The primary aspect of the mechanism that these mathematical models simulate is receptor mediated signaling, all of which work through the corticosteroid receptor.

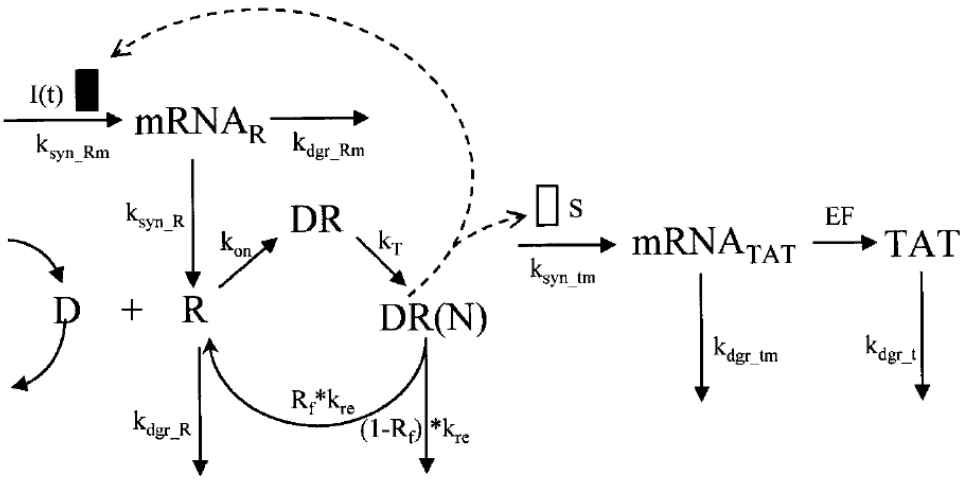


Figure 1: A schematic which shows the primary components which underlie the fifth-generation model of corticosteroid activity. This model was developed to replicate the activities of Tyrosine Amino Transferase (TAT), a marker gene that was selected from the mRNA microarray. Figure was obtained from [10].

However, while such work has provided a great deal of insight into the underlying systems, specific limitations still exist in the current model. The fifth generation PK/PD model was developed to describe the response of an organism to a one-time injection of corticosteroids. Predictions of this model were then validated through the use of a double-dosing of corticosteroid[3], and the initial tolerance hypothesis seemed to be supported by the experiment. However, the model failed to predict the response of the organism in response to a continuous infusion of corticosteroid. The contradiction that existed between the model predictions and the observed responses lay in prediction that the effect of corticosteroids upon the system should be non-existent after 24 hours, whereas physiological indicators such as muscle wasting suggested that corticosteroids were having a prolonged effect upon the system[12]. This is because the fifth generation model is predicated upon the hypothesis that the corticosteroid receptor should show no activity after a short time period due to a tolerance mechanism[10] due to degradation of the receptor. Such mechanisms are not able to explain the sustained metabolic effects of corticosteroids such as a constant loss in muscle mass when corticosteroids are infused at a constant rate rather than introduced via a bolus injection. Finding a model which can explain these observations would go a long way in enabling us to determine how to target the inflammatory pathways independently of the metabolic pathways, thus increasing the therapeutic effectiveness of corticosteroids. Furthermore, in addition to the direct question involving corticosteroids, we are also interested in creating a generalizable framework which can be used to tackle a wide variety systems, not just those associated with the corticosteroid response.

Taking the previous model into account, our starting point will be the assumption that the primary response of corticosteroids occur through transcriptional regulation. Thus, it is our hypothesis that whatever changes are wrought by an administration of corticosteroids, their

activity can be seen via changes in gene expression. However, because we do not know *a priori* which genes are affected by an administration of corticosteroids, we have elected to utilize a high throughput dataset in which the temporal gene expression profile of thousands of genes have been captured. Because the arrays themselves do not represent a comprehensive set of genes which are expressed, of primal concern to us is the extraction of patterns indicative of corticosteroid activity rather than the identification of every gene which response to corticosteroids. These patterns can then be used as a feature for future selection if the need arises. After a set of candidate genes have been identified, the next step is to determine the underlying mechanism which governs the responses of these genes. To perform the identification of this underlying mechanism, we will approach the problem in three different ways to determine whether it is possible to obtain mechanistic insights:

1. The identification of hypothetical regulatory elements
2. Linking the activity of the regulators of mRNA
3. Linking the activity of the effectors of corticosteroid activity to drug administration

Finally, after these mechanistic insights have been linked, we will propose a model of corticosteroid activity, and determine whether it is able to reflect the observed response to corticosteroid administration.

The microarray data which forms the foundation of this dissertation will be based upon the mRNA isolated from the rat liver. While corticosteroids have been shown to have a significant effect upon multiple tissue systems such as, but not limited to the liver, kidney, bone, muscle, and fat[13], we hypothesize that because the liver plays an important role in the detoxification and elimination of drugs[14], as well as its central role in both metabolism and inflammation, it

offers the simplest way for us to determine the possible mechanisms of how a tissue responds to an administration of corticosteroids. Though we focus primarily upon the analysis of the liver, we do not discount the fact that there may be significant interactions between the different tissue systems through the alteration of metabolites such as glucose and glutamine. However, before increasing the complexity of the system and analyzing the systemic effect of corticosteroids, it is important to first establish a mechanism by which one individual tissue has been affected.

The two primary gene expression datasets which we will be utilizing involve two methods for delivering corticosteroids into a rat. The first method is a bolus injection of 50 mg/kg of corticosteroids into a rat animal model. In this animal model, the rat has been adrenalectomized so that the only source of corticosteroids is the injection of corticosteroids. In the second dataset, adrenalectomized rats were infused with corticosteroids at a rate of .1 mg/kg/hr. The removal of the adrenal gland allows the levels of corticosteroid in circulation to be accurately controlled as an experimental parameter without having to compensate for the adjustments made by the adrenal gland in response to high levels of administered corticosteroids[15].

The structure of these inputs mimics the inputs normally used for system identification such as the impulse function and a step function[16-18]. The impulse function corresponds to the bolus injection of corticosteroids, and can be used to obtain the transfer function from linear systems. Step functions can be used to obtain a rough estimate of various system factors such as time lag, gain, and delays. Therefore, the experimental datasets which have a solid foundation to be used for model building.

In these experiments, a population of rats is first temporally synchronized through fasting and in the synchronization of the rat's light/dark cycle through environmental condition. After the synchronization steps, the rats are administered corticosteroids, and at predetermined time

points, a small subsets of rats are killed to provide technical replicates. This destructive sampling essentially treats the different animal replicates as coming from a single “giant” rat[19]. After the animals are sacrificed, the liver is excised, the mRNA isolated, and then quantified via Affymetrix microarrays. Thus, for each experiment, we have a set of microarrays which will span the experimental time interval with 3-4 replicates per time point. These time points can then be treated as samples obtained from a dynamic response. It is then our task to first determine the underlying patterns within the data, and then derive a model which explains these underlying patterns. All of these datasets can be obtained from the Gene Expression Omnibus (GEO) database. The dataset corresponding to the bolus injection of corticosteroids is present under the accession number GDS253. In this dissertation, it shall be referred to as the acute dataset. The infusion of corticosteroids, which will be referred to as the chronic dataset is present under the accession number GDS972.

Equation Chapter (Next) Section 1Data Validation

While not specifically part of the two steps outlined through our investigation of emergence, data validation is still an important part of our evaluation. While the use of high throughput methods such as microarrays attempt to let the data infer the underlying mechanism rather than validate the underlying mechanism, it is still important to determine whether the data itself represents a coherent response[20]. However, unlike most methods which attempt to validate the accuracy of mRNA microarrays with respect to other measurement techniques such as RT-PCR[21], we seek to validate that the experiment as designed will show significant coherent changes in the system in response to an administration of corticosteroids. In our specific case, it is still undetermined whether corticosteroids themselves will affect every tissue system in the body, or whether the sampling strategies which have been utilized are sufficient in capturing the underlying response of the system. Therefore, before any more involved work is undertaken, it is imperative that that one determines whether the data itself has captured meaningful dynamics. Because the primary goal of this dissertation is the study of how corticosteroids impact the biological system, the focus is upon transient dynamics rather than permanent alterations in a cell's state as in the case in comparing the difference between wild-type and mutant strains[22], or between cancerous vs. non cancerous cells. Because of the need to capture these transient dynamics, the issue of sampling becomes an important one[23]. If the system is incorrectly sampled, fast dynamics may not be accurately captured **Figure 2.**

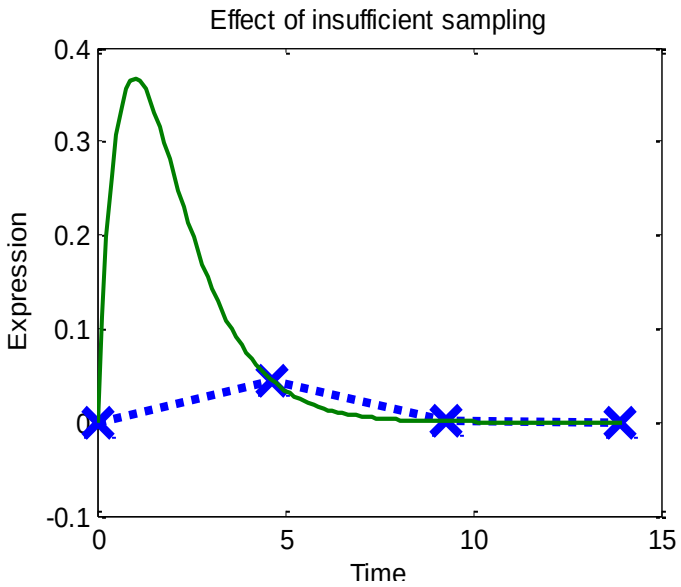


Figure 2: In the simulated dynamic with a rapid early response, and a slow decay back to baseline, insufficient sampling may lead to a sub-optimal reconstruction of the signal which may lead to incorrect assumptions about the dynamics of the system

Basic sampling theory indicates that to accurately capture any dynamic response it requires that the sampling occur at twice the natural frequency. This is known as the Nyquist sampling theorem[24]. Essentially, the Nyquist sampling theorem states that a periodic sine wave can be reproduced so long as more than 2 points per cycle has been recorded. Furthermore, most signals can be represented as the summation of sine and cosine curves via a Fourier transfer[16]. Thus, in the general case, a signal should be signaled at twice the rate of its highest frequency component. However, in the context of biological systems, such sampling requirements are often not met. One of the most obvious reasons for not sampling at the Nyquist frequency is the simple fact that the fastest dynamic needs to be known before the experiment can be carried out. Because this information is not always present, researchers can only make an estimate as to how fast the quickest dynamics are. Furthermore, given the fact that the temporal scales of the dynamics are unknown, it is very possible that the dynamics which we wish to capture have been missed in their entirety.

From the perspective of electrical engineering, the simple answer to this dilemma is to oversample the system such that a comfortable margin is obtained. However, unlike in electrical engineering, the penalties for overestimating the frequency of the most rapid dynamics within a given system are quite significant. In our experimental context, each time sample comprises of multiple animal replicates. Each animal represents a significant investment in time and money. Therefore, with the fastest gene expression dynamic is on the order of minutes[25], it would be impractical to sample at such a rate[23]. Rather, researchers have divided the dynamics into a rapid early response and a slow late response[26]. However, because this information relies specifically upon researcher intuition, it is still imperative that the data still be validated, i.e. determine whether the dynamics which have been captured can be used to rationally assess a given system[27].

The evaluation of our dataset is two-fold, first to evaluate whether the dynamics can be used in a model building exercise, and secondly whether the dynamics can be used to identify significantly activated system in the biological organism. To tackle the first question, we need to validate whether the dynamics captured by the array show significant non-random behavior. The challenge associated with this task is to determine whether the array itself represents some non-random dynamic behavior rather than whether a given gene shows significant non-random behavior. Assessing the entire array is important because of the initial hypothesis that genes which respond to some external perturbation show significant coordination of activity. Thus, we seek to validate whether this is true in a given array set before conducting further evaluation.

Determining the Validity of the Dataset

Before, this can be accomplished; we must first define the null hypothesis, i.e. the characteristic of data which shows no significant activity or coordination. Because of the interest in obtaining models which can describe the underlying dynamics associated with the drug administration, we are concerned whether a set of gene expression profiles can be represented via some mathematical model. One possible generalization of these signals is the Auto-regressive Model (AR)[28], in which the signal is defined as in (1.1) where x_n represents the signal at time point n whose value is a linear combination of previous time points, with the scalar α**.** In this model, one assumes that the value of the function at time point N, is defined by a linear combination of the values at p previous time points. This model is attractive because it represents an analogue to a set of set of linear differential equations.

$$x_n = \sum_{i=1}^{n-1} \alpha_i * x_i$$

(1.1)

Thus if one were to take a set of differential equations as the basic mathematical model describing gene expression, a reasonable question is whether the recorded gene expression profiles reflect such behavior. By taking the auto-regressive model, one essentially assumes that the value for a gene expression profile at time *t* is dependent upon previous time points. To determine whether a gene expression profile shows this relationship, an autocorrelation function defined in (1.2) will be used. In this formulation R_{ff} defines the autocorrelation for a specific offset denoted τ. This function consists of the integral between a signal, and an offset version of itself. Because of this offset, it is able to determine to a limited degree the relationship a given time point has with previous time points.

$$R_{ff}(\tau) = \int_{\infty}^{\infty} f(t+\tau) * f(t) dt \qquad (1.2)$$

The auto-correlation function shifts the signal in relationship to itself. If the signal at time t has no relationship to previous time points, then the magnitude of the auto-correlation function should be close to zero. If there is a relationship, then it should show some intermediate value. This is illustrated in **Figure 3**, where the auto-correlation of a long random signal is taken.

The results of running the auto-correlation function upon the synthetic dataset shows that at lag = 0, the correlation is perfect. The lag in this case is τ or the offset applied to the signal when it is being compared to itself. This is not surprising because a signal will correlate perfectly with itself. However, when the lags are increased, the random dataset goes immediately to a value close to zero and fluctuate around this value. Therefore, the region of interest to us in our auto-correlation evaluation is the tail region i.e. lags ≠ 0. To account for the different number of genes, and the time length of the individual signals, we will be focusing primarily upon the region

corresponding to lag = 1 to lag = T, where T is the number of time points which have been recorded.

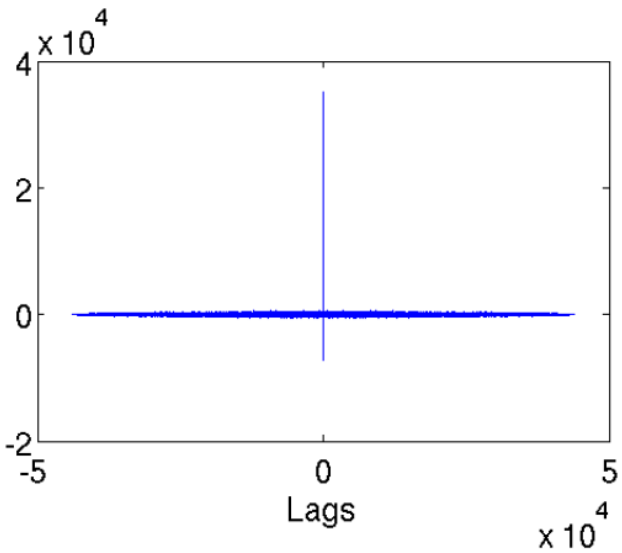


Figure 3: The autocorrelation function for randomly generated data. Note that at τ = 0, that the signals are perfectly correlated, but the correlation falls to a very low level, at τ≠ 0, due to the lack of relationship between adjacent time points.

The notable property of using the auto-correlation upon microarray data is that it is sensitive to any sort of order imposed upon the dataset. In this algorithm, the goal is to detect the presence of order within each gene expression profile hinting at the fact that coherent changes in gene expression have been captured by the experimental design. However, if order was imposed upon this system through by ordering the genes in some way, such as by ordering the genes by the value of their expression at the first time point, then the second time point, etc. it is possible to obtain a significantly different behavior in the auto-correlation function as shown in **Figure 4**. Therefore, any order within the system whether present due to the relationship between adjacent time points, or order imposed artificially can be detected by the auto-correlation function.

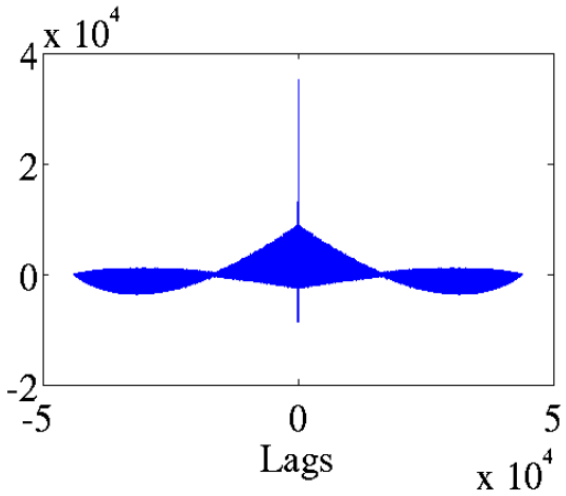


Figure 4: Imposing order upon the dataset through a simple sort, greatly changes the dynamics of the auto-correlation function. This is because there is a loss of randomness from the sorting operation

The dynamics of this correlation coefficient can vary either through a difference in the mean, or a difference in the standard deviation or both. As seen in **Figure 4** the dataset which corresponds to the random data and the result that corresponds to the sorted data appear to have significantly different trends. Therefore, if it were possible to quantify these trends, it may be possible to obtain a metric that quantifies how dissimilar a given dataset is to the reference null dataset, with the expectation that the closer a dataset is to the null hypothesis the less meaningful such a data set is.

To quantify the difference between the null dataset and our dataset of interest we have selected the f-test. This is because given the auto-correlation of random data at lag ≠ 0; there should be very little fluctuation about zero, whereas in informative data, we expect to see greater fluctuations because of the slower loss of correlation due to relationships between the time points. The f-test is defined in (1.3) where s_{xx} represents the covariance in X, and s_{yy} represents the covariance in Y. Again as in the f-test, X represents the auto-correlation of the randomly generated dataset and Y represents the autocorrelation of the microarray dataset**.**

$$f = \frac{s_{xx}^2}{s_{yy}^2} \tag{1.3}$$

Previously it has been shown that the auto-correlation is sensitive to any order imposed upon the system. Given that the microarray itself is an artificial construct, there is the possibility that there would exist significant structure within the data as presented, and thus alter the response of the auto-correlation function. Therefore, to eliminate this possibility, we will first randomly permute the rows in a given dataset before concatenating the gene expression profiles into a single NTx1 vector. This eliminates all possible prior structures within the data expect for the

relationship between the gene expression level of a given gene at time T vs. the gene expression at T-1.

Thus the steps of the algorithm are as follows.

1. Concatenate the response of the separate genes from an NxT dataset into an NTx1 dataset
2. Calculate the auto-correlation function
3. Compare the auto-correlation function of the real dataset vs. a randomly generated dataset at |lag| < 100 with the same number of genes and time points via the f-test, and obtain a p-value describing the statistical significance of the data.

Assessing the Consequence Validated Data

To assess whether temporal gene expression datasets which corresponding to a high p-value are more meaningful than temporal gene expression data with a low p-value, we will exploit an unsupervised classification techniques known as clustering. The hypothesis behind using clustering is that since they work primarily off of the temporal evolution of expression values, the more reliable a set of gene expression profiles, the more reliable a set of genes would have been assigned to their respective functions[29]. This hypothesis is termed “guilt-by-association.” These biological functions are known as gene ontologies and are present in databases such as gene-ontology.org[30, 31], but also as annotations on most commercial microarray platforms such as the Affymetrix arrays which most of our data is presented in.

A brief review of various clustering techniques is given in[32]**.** The primary hypothesis behind utilizing clustering algorithms is that system with similar responses ought to have some intrinsic similarity. In the field of biology, it is hypothesized that genes which similar responses to a given

input stimuli do so because they are part of a coherent mechanism that responses to perturbations away from homeostasis. Thus, while there exist multiple ways of assess the success of a given clustering operation, such as calculating the information content of a given clustering or the inter/intra cluster distances[33], we have elected to utilize the biological hypothesis because of the relationship it will have with later analyses. Therefore, if the individual gene expression profiles were not properly captured and the dynamics are essentially random, clustering algorithms will lump the gene expression profiles in an essentially random manner. Thus, any functional patterns within the data should not exist. However, if the temporal gene expression profiles have been accurately captured, then similar functionalities should be evident. Thus, in an informative dataset, a given clustering result should yield many genes which have similar functionality. This concept can be used to evaluate the quality of a given clustering algorithm if the dataset remains constant[34], and likewise be used to evaluate the quality of a given dataset if the algorithm remains constant.

To assess the statistical over-representation of a given biological functionality in a given dataset is termed the enrichment, and is measured via the hypergeometric distribution(1.4). The hypergeometric distribution essentially calculates the probability a given subset of functions will be chosen at random from a larger population of possible functions[34].

$$P = 1 - \sum_{k=1}^{n} \frac{\binom{k}{i}\binom{m-k}{N-i}}{\binom{m}{N}};$$

n = number of times the ontology appears in a given cluster
i = number of genes in a given cluster (1.4)
N = total number of genes
m = number of times the ontology appears in the dataset

One of the issues which we need to address with clustering is the identification of a concrete number of clusters. Given a single dataset, the methods by which to determine the optimal parameters associated with a given clustering algorithm is still an active area of research. Therefore, rather than determine the optimal number of clusters, the quality of the dataset will be judged by how enriched it is over a continuum of different cluster numbers. Thus, if the proposed metric is successful, then what should be evident is that if a dataset is more informative, then it should be more enriched for a given number of clusters. To perform this operation, we will be taking the continuum of clusters numbers from 2 clusters to 19 clusters. The clustering operation will utilize cluto[35], a widely used clustering package for clustering time series data.

Given the existence of a second method for identifying the informativeness of a given dataset, one may question the need to formulate another method as was done here. The key difference between the methods is that the use of gene enrichment to assess the quality of a given dataset requires a significant amount of external information in the form of gene ontologies. Thus, for one to quantify datasets through ontology enrichment, it requires the use of well studied systems in which the roles of many genes have already been elucidated. Furthermore, because the use of ontologies requires external information, such methods are not applicable to all temporal data such metabolic fluxes, in which there isn't necessarily a link between co-expression and co-functionality. Thus, we are proposing a method which is generalizable to all temporal data, and comparing it to a technique that is limited to well studied systems which can be captured via mRNA gene expression. Thus, the proposed method essentially functions as a simple standalone method to quantify a set of temporal signals.

Test Data

Along with our two corticosteroid datasets, we will utilize a randomly dataset to establish the performance of the algorithm upon a null dataset. This null dataset will be used to show that if the algorithm does not return any significant difference between the two datasets there will also be very little improvement in the gene ontology enrichment after the clustering step, thus negating the ability for clustering type approaches to identify co-regulation or co-functionality. Secondly, we will utilize an additional dataset present under the GEO omnibus ID GDS802[36] to establish whether any significant trends appear within the method. The third dataset to be used for this evaluation is a short term burn dataset which of mRNA gene expression data obtained from the liver after a rat animal model had been exposed to a full skin thickness burn consisting of 30% of the animal's skin area.

Dataset	p-value
Random Dataset	0.99
Acute Administration of Corticosteroids	0
Chronic Infusion of Corticosteroids	$2.7x10^{-4}$
Burn Dataset	0.96

Table 1: The F-test values for the four different datasets. The prediction is that the Acute corticosteroid dataset will be the most informative whereas the Burn Dataset will be the least informative

Benchmark Run

The data quality metric which corresponded to the four datasets are given in **Table 1**. Looking at the first 100 offsets or lags, we can see that there appears to be an inherent structure present within the auto-correlation tail in two of the datasets **Figure 5**. These datasets are the two corticosteroid datasets. In these datasets, we see two features. The first feature appears to be the periodic spike of the auto-correlation function, and the second feature appears to be a highly non-random structure interspersed within the spike train. The first feature suggests that significant portions of the data are well correlated i.e. there appears to be a dominant pattern within the data. The second feature suggests that the individual dataset can be modeled by an auto-regressive function, because there is some relationship between a gene expression at time N and previous time points. The dataset that corresponds to the burn dataset on the other hand does not appear to have any significant structure and appears to respond similarly as the theoretical dataset which corresponds to the null hypothesis.

From this analysis we hypothesize the fact that the corticosteroid datasets appear to be more informative than the burn dataset. Both of these datasets show a significant amount of internal structure as well as a significant amount of highly correlated gene expression profiles. To validate this we had elected to run clustering and gene ontology enrichment. The results of gene ontology enrichment are presented in **Figure 6.** What is evident in this figure is the fact that the datasets which were indicated as informative were more enriched for a given number of clusters than the datasets which were indicated as less informative. The conclusion which is drawn from this is that the gene expression profiles can be more accurately grouped if the data has been well measured.

Having established that the proposed algorithm agrees with the results of a more traditional approach, we can then utilize this algorithm to assess all of the high throughput datasets that will be used for the rest of this dissertation. Thus we have established the fact that our datasets are indeed informative. Thus having established the fact that our datasets are good, we can begin to conduct our examination of the system from previously obtained data.

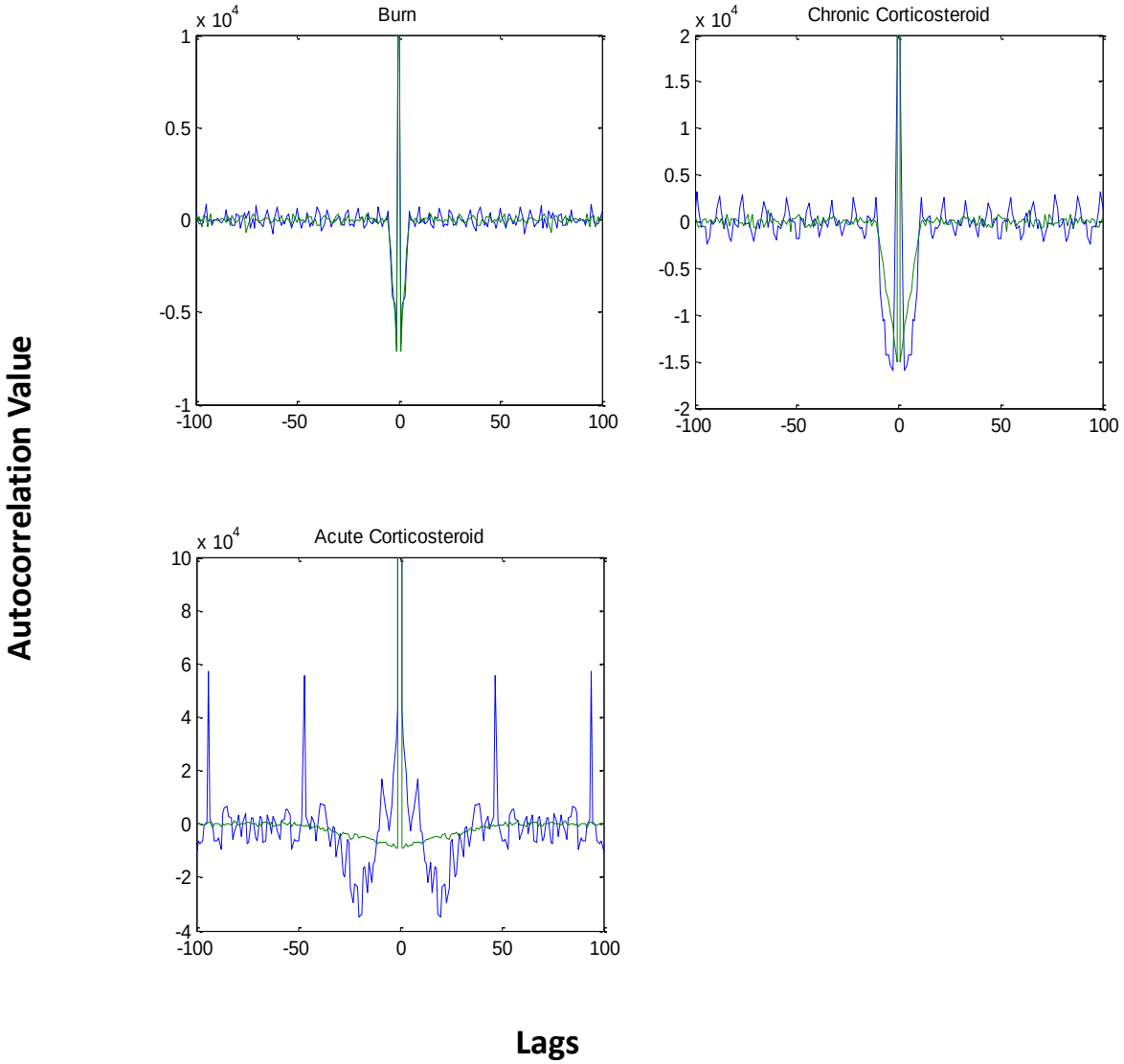


Figure 5: The autocorrelation function of the three real datasets. The green line represents the null response, whereas the blue line represents the response of the dataset. The features of interest for us are the presence of the periodic spike trains in the chronic and acute corticosteroid data, and the slow oscillations around zero. The periodic spike train denotes a large number of the genes show significant co-expression, whereas the slow oscillations around zero suggest that there are significant relationships between adjacent time points. This would indicate that sufficient sampling has occurred.

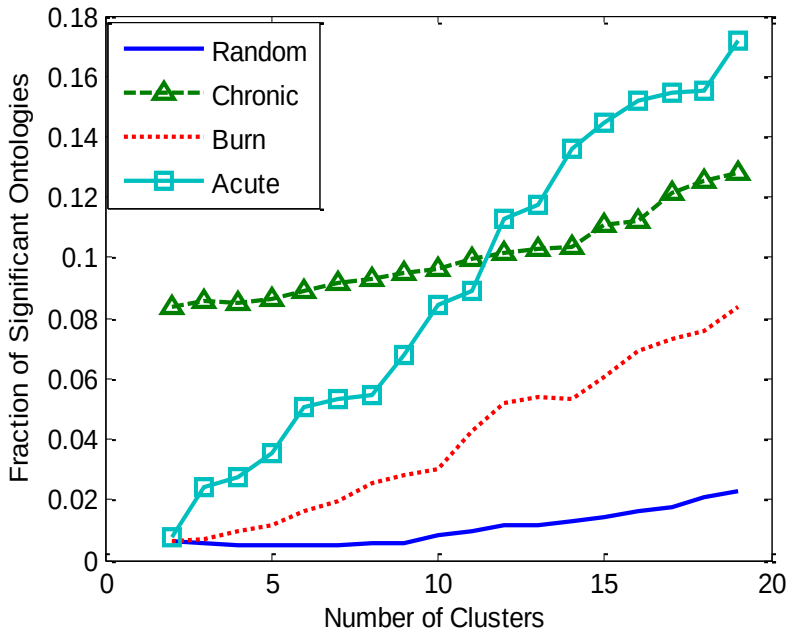


Figure 6: The number of enriched ontologies as a function of cluster number. The specific feature of interest is the fact that the datasets which show low p-value (high significance) under our metric appear to have significantly larger faction of ontologies which are enriched as compared to either a randomly generated dataset, or one with a low p-value (burn dataset)

Discussion

One of the primary motivations behind our Systems Biology analysis is that all of the algorithms should be unsupervised. This means that there will be no gold standard by which we can calibrate the results of the algorithm. This leads to a confounding problem in that so long as the data is in the proper format, an unsupervised algorithm will return a result. For instance a clustering algorithm such as k-means clustering will be able to cluster any kind of data, but not in all cases will the clustering result be meaningful. This presents some difficulty because the researcher cannot be guaranteed that the results are a consequence of the biological function which has manifested itself as a coherent pattern within the data, or if the results are an artifact of the algorithm itself. Therefore, it is important for us to be able to establish *a priori* that the data is indeed good.

In the results we see that the gene enrichment analysis corresponds closely with the p-values associated with the f-test, and therefore we can establish a general trend, linking the two data metric qualities. One of the results seems to suggest that one of the datasets is meaningless. Specifically, it was found that the dataset corresponding to the burn dataset did not significantly differ from that of a randomized dataset. However, we must qualify our classification of meaningless. While our algorithm has suggested that the burn dataset was not informative, it is only non-informative when one tries to process the data under the hypothesis that co-expression implies co-regulation or co-functionality. On the other hand, it may be possible that the burn dataset may be able to identify the importance of specific genes through differential expression. But, by focusing only upon differential expression over time, we are unable to exploit the temporal nature of the data, specifically the identification of coordination between the disparate systems, or the construction of dynamic models which describe its activity. However, unlike the dataset which corresponded to a burn, the corticosteroid datasets all exhibit

significant amounts of information. Thus, it appears to us that a clustering approach may be a viable method for analyzing the data rather than simply looking for whether a gene is up/down regulated.

While it is arguable whether our algorithm is in fact needed due to an alternative technique which is able to yield similar results, our algorithm has two primary benefits. The first obvious benefit of our algorithm is speed. Establishing the suitability of a given dataset for analysis required running a clustering algorithm multiple times, and then mining a database for a given ontology. Depending on the size of the datasets, the two operations can be prohibitive. In our speed tests, we have found that our algorithm ran in less time than it took to generate all of the clusters for ontological evaluations. Secondly, our algorithm was able to establish the presence of relationships between adjacent time points in the data. This allows us to determine whether automated model building operations are a viable alternative because sufficient time samples have been taken, something that the standard approaches cannot accomplish.

Conclusion

It was important for us to establish that our datasets were meaningful. In light of the automated analysis which will be presented later in this dissertation, it is important to establish the primary components of our analysis will return meaningful information, without requiring us to qualify our results with various caveats due to ambiguities within our data. Thus, the analysis which we have performed upon our datasets has established that we can run clustering algorithms to identify the presence of important functionality within our data. Secondly, it also suggests that we ought to be able to utilize the data to build dynamic models which describe the response of the system.

Because the method was developed in response to the question as to whether the dataset was meaningful in the context of whether sufficient temporal sampling had occurred such that relevant information could be extracted, one extension which we are interested in is whether it is possible to obtain a metric that would inform how many additional sampling points are needed to convert a non-informative dataset into an informative dataset. If this could be accomplished, then it would give researchers a systematic method for sampling dynamic systems without the risk of over or under sampling. If this metric existed, then we envision a researcher running a specific experiment, evaluating this metric, and then from the metric determine how many additional time points are necessary.

Equation Chapter (Next) Section 1Analysis of Gene Expression

After having established the validity of the datasets which will form the basis of our analysis, it is then possible to begin the analysis. There currently exist numerous methods which are used for the analysis of microarray data. They usually fall into two separate categories, selection algorithms[37] and clustering algorithms[38]. Given the fact that microarrays measure a large and standardized set of genes, the logical first step is to identify those genes which respond to the underlying treatment. This hypothesis forms the basis of the utilization of selection algorithms. If in addition, the data is a time series as in our case, the clustering step will be carried out to mine for significant patterns within the data. These significant patterns can later be used to identify major processes which are related to the phenomenon being studied, creation of dynamic models, and finally the identification of common regulatory processes.

Rather than relying upon pre-existing methods, we have elected to formulate a new method to perform both the selection as well as the clustering of the data. The primary motivation for the creation of a new algorithm is our hypotheses that by focusing primarily upon differential expression, the temporal aspect of our datasets are not fully exploited[39]. Most gene selection algorithms are focused upon the selection of genes which have been shown to be differentially expressed above a certain statistical threshold. For instance, one common method for selecting temporal gene expression profiles is the use of the ANOVA[40]. This test is a generalization of the t-test[41] over multiple time points or conditions. Genes that are selected via ANOVA are those which have shown statistically significant differential expression over the experimental time course given a certain number of replicates, and essentially work to identify whether the organism can be shown to reach a new state over the duration of the experiment. Other gene selection methods such as SAM[42] and PDNN[43] essentially perform the same task. A more

thorough review over the different methods is presented in [44, 45]. The primary difference between these different methods lies in alternative ways of assessing the level of differential expression, or calculating the significance level associated with the differential expression.

The primary assumption behind the use of differential expression is that if there are no significant perturbations to a system, then the underlying gene expression ought to be static. However, this contradicts the fact that homeostasis is a dynamic process as seen in the inherent dynamics associated with an organism at rest[46]. Because the organism itself is responding to intrinsic changes, it should not be surprising that genes will change their expression level. Because of the underlying variance associated with the dynamic nature of homeostasis, it would be possible that given a large enough number of replicates that even in an unperturbed system there could still be genes that are differentially expressed. Thus the use of differential expression metrics such as ANOVA or SAM is essentially assessing how accurately a given gene expression profile has been measured, and thus more dependent upon the number of replicates than the underlying biological significance of a gene[47].

We instead propose an alternative hypothesis for the selection of informative genes. Rather than utilize a selection criteria that is designed to compensate for the limitations associated with the measurement processes, our alternative selection criteria is based upon a hypothesis as to what biologically significant genes ought to show. Our hypothesis is that genes do not respond to external perturbations in isolation, but are rather in coherent groups. Thus, if a gene is related to a given biological response, it ought to be correlated with a large set of other genes.

Because our selection criterion hypothesizes that genes that are important will be correlated to a large number of other genes, it requires the use of a clustering algorithm, preferably one which is able to divide genes up into a large number of different clusters. While there exist a large

number of clustering algorithms that could be used we have elected to adapt the HOT SAX formulism for the purposes of clustering[48]. The HOT SAX formulation is attractive for various reasons.

1. The HOT SAX algorithm is fast and scales linearly
2. The HOT SAX algorithm is deterministic
3. The intermediate results of the HOT SAX algorithm provides important intermediate results that provide insights into the dataset

After the clustering step, we then have a large number of clusters. At this point, the task is to determine which of these clusters are actually relevant to the underlying response of the system. In a related method by Bar Joseph et al.[49], they have elected to look only at the population of a given cluster after running a similar fine grained clustering algorithm. Their hypothesis was that given a randomly generated dataset, the population of a specific cluster can be modeled via an underlying distribution, and thus the clusters which had more genes than expected were selected as being significant.

We have elected to utilize a different metric instead of relying upon the population of a cluster. Our selection criterion instead assumes that given a base state, a population of genes will conform to some underlying distribution of expression values, and as the organism responds in a coordinated manner, this distribution will change. By looking at the population distribution, we can minimize the effect of the dynamic fluctuations of individual genes. Only when there is a large coordinated change among all of the selected genes will the underlying distribution change. However, rather than select clusters that deviate past a certain threshold, we will be selecting clustering which maximize the observable change in the distribution of expression

values. This selection thereby allows us to select a set of candidate genes which characterize the system. Because of the integrative selection step and the utilization of HOT SAX, we have termed this algorithm SLINGSHOTS for the (SeLection of Informative Genes via Symbolic Hashing of Time Series)[50].

Hash Based Clustering

HOT SAX was initially developed to find short recurrent patterns within a long time series[48]. It does this by taking a sliding window and converting the signal within the window into an integer. This allows one to efficiently scan through a given signal looking for over-represented motifs rather than having to conduct a One Against All (OAA) Comparison for each window. We have taken this formulism and adapted to the clustering of relatively short time series. Thus, rather than finding patterns within a single time signal, we are looking for patterns over multiple sets of time signals.

The HOT SAX clustering method is a four step process.

1. Normalize the Gene Expression Profiles via the z-score
2. Piecewise Average the Gene Expression profiles if necessary
3. Quantize the expression profile with equi-probable breakpoints
4. Convert the quantized form of the signal into a single integer

The first step which comprises up of normalizing the gene expression profile is performed so that genes with similar shapes, though differences in magnitude will be grouped together. The Equation for z-score normalization is defined via (2.1) where Y represents the underlying signal, <Y> represents the mean of the signal and σ(Y) represents the standard deviation of the signal.

This essentially converts all of the signals such that they have a mean of zero and a standard deviation of one. Performing this z-score normalization allows us to define a consistent set of breakpoints with which to quantize each temporal gene expression profile.

$$\widehat{Y}_i(t) = \frac{Y_i(t) - \langle Y_i(t) \rangle}{\sigma\left(Y_i(t)\right)}, \quad \forall i \tag{2.1}$$

The second step is optional, and is dependent upon the length of the data. As a general rule of thumb, signals with greater than 11 time points will be piecewise averaged to convert the signal to a shorter time series. This accomplishes two things, the first consequence is that signals will have been low pass filtered and the second is that the shorter signal becomes more numerically tractable for the third step. The consequence of the low pass filtering is that many fast responses may be diminished or lost. Because of the low pass filtering effect of the piecewise averaging; any piecewise averaging should be conducted such that the reduction in the signal length should be kept to a minimum. Therefore, if one were given a signal of length 12, the piecewise averaging should average two adjacent points leading to a new signal with a length of six, rather than averaging 3 adjacent points leading to a signal of length four. Shortening the signal any more would lead to a greater loss of information. The number of adjacent points to be used in piecewise averaging is one of the two parameters that need to be selected by the researcher for this algorithm.

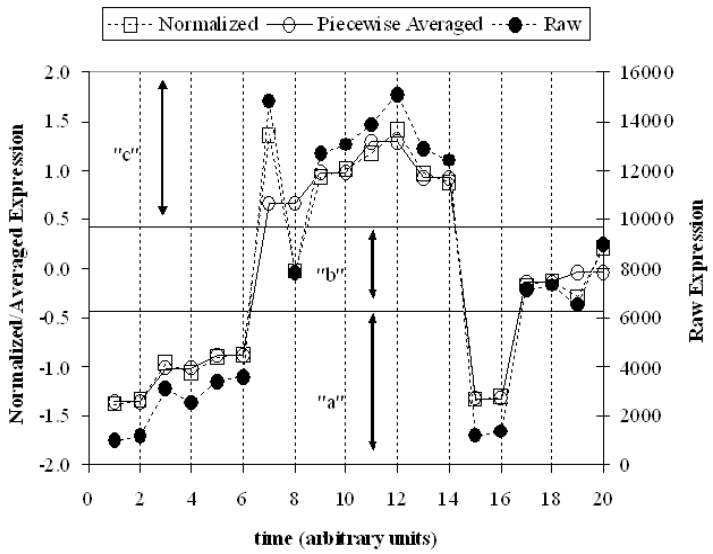


Figure 7: A schematic denoting the process of converting a temporal signal into a string of symbols

The third step involves first identifying the number of partitions used to represent a symbol. Therefore, if three partitions are used to represent the signal, the normalized analog signal will be converted into a string of letter “a”, “b” and “c”. Given four partitions, the signal will be converted into a string of “a”, “b”, “c” and “d”. With more partition, the alphabet size that describes the signal will increase accordingly. However, unlike the standard partitioning with equi-distant partitions, we will instead be using equiprobable partitions. The use of the equiprobable partition means that if a signal has been drawn from an N(0,1) distribution i.e. random, the appearance of a given symbol, “a”, “b”, “c” will occur with equal probability. Because of this equiprobable distribution of symbols, we can use this fact to evaluate certain properties of our dataset.

In order to identify where the breakpoints should be placed, we assume that a random (null) signal conforms to a Gaussian distribution. We therefore need to select breakpoints such that the area under the Gaussian curve is equal. These breakpoints can be obtained via the use of tables found in standard statistics book, or through the use of (2.2). In (2.2), the breakpoints are obtained by solving for x where n is the number of breakpoints z is an index variable that goes from 2 to n, and erf is the error function. Theses breakpoints are used to quantize the signal into a string of symbols as shown in **Figure 7**. After the gene expression profiles have been converted into a string of symbols, we then convert this string into a single integer. The hypothesis is that genes that hash to the same integer will have the same symbolic representation, and thus have similar gene expression profile. The conversion from a set of symbols into an integer is accomplished by (2.3) and is analogous to converting a base n number into base 10. In (2.3) AB is the size of the alphabet, w is the length of the signal, C is the character representation of a given signal at time point j, and H is the hash value for gene i.

$$\frac{n}{z-1} = \frac{1}{2}\left[1 + erf\left(\frac{x}{\sqrt{2}}\right)\right] \tag{2.2}$$

$$H_i = 1 + \sum_{j=1}^{w} [ord[C_i(j)] - 1] \cdot card(AB)^{w-j} \tag{2.3}$$

Parameter Selection

In order to utilize HOT SAX, two parameters must be chosen. W denotes the number of adjacent time points to piecewise average and alpha denotes the number of breakpoints to use. While the selection of w can be made by looking at the number of time points in a given dataset, the selection of alpha is less obvious.

Because the HOT SAX algorithm utilizes equiprobable breakpoints, a randomly generated signal will be assigned hash values from 1 to AB^t with equal probability. Given a population of randomly generated signals, the probability that N signals will hash to the same value can be modeled by the Poisson distribution[51]. In the case where AB^t is greater or equal to the number of genes being clustered, this can be approximated by the exponential distribution. However, in a real biological experiment, one expects that the temporal gene response will not be randomly generated, and manifest significant coordination between the different genes as the organism responds to challenges to homeostasis. Thus, if there is some inherent underlying structure within the data, the probability that N genes will be assigned the same hash value should not correspond to the exponential distribution. Therefore, AB should be selected such that there is an observed deviation from this hypothetical exponential distribution. The correlation between our distribution of population values and the best fit exponential distribution will be quantified by the R^2 correlation coefficient.

To illustrate the response of the HOT SAX algorithm to a randomly generated population of gene expression profiles, we have constructed a synthetic null dataset where each data point is drawn from the N(0,1) distribution. Looking at the distribution of cluster populations it is apparent that this dataset appears to correspond to the hypothetical exponential distribution. Furthermore, utilizing different parameters does not change the response of the random dataset. Our experimental datasets on the other hand show a different response. Running the hashing operation upon one of our experimental datasets, we see the following response. Due to the fact that there is some coordination which occurs in this dataset, there exists at least one AB which shows a significant deviation from the hypothetical exponential distribution.

For the current implementation of the algorithm, we have calculated breakpoints for 3-5 breakpoints. Thus to obtain the most optimal number of breakpoints, the hashing operation will be run parametrically for these three breakpoints, and the one which yields the slowest R^2 correlation will be selected as the optimal value for AB.

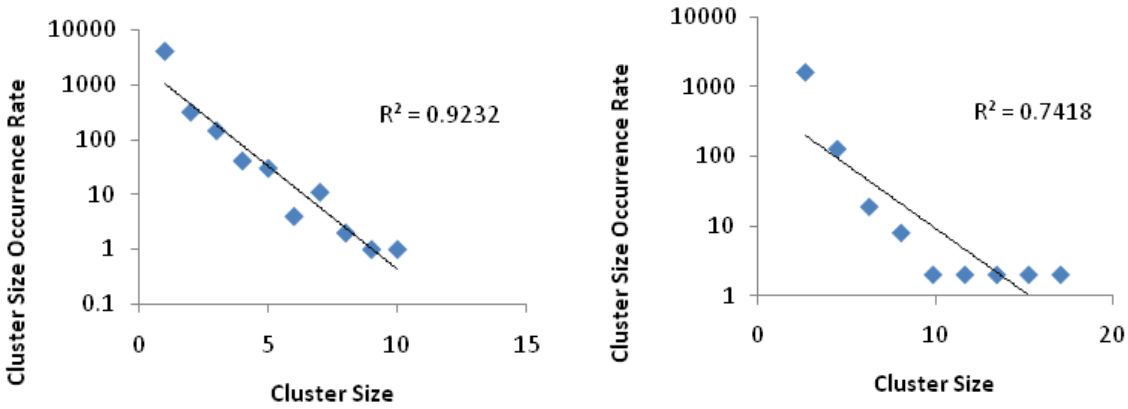


Figure 8: The population response of corresponding to different datasets. The dataset on the left represents the result corresponding to our null dataset. The dataset on the right represents an informative dataset because of its deviation from an exponential distribution which would be characteristic of a synthetic null dataset. For a given dataset, the correlation will be evaluated for the different alphabet sizes (3,4,5), and the alphabet associated with the lowest R^2 correlation will be chosen as the optimal parameter.

To quantify how far the population distribution of hash values for a given data set is from the hypothetical exponential response, we take the R^2 correlation between the population distribution and the best fitted exponential distribution. This is then compared to the R^2 correlation of a large set of random trials from a null synthetic dataset that has the same number of time points and the same number of signals. These random trials than allow us to report a confidence level that a given dataset is not synthetic and therefore informative. The results of this analysis should correspond to the results which the assessment of data quality present in the previous chapter. However, it is important to note that the two metrics are not identical. This is because the metric proposed in the previous chapter is not only able to discern the relationship between the different gene expression profiles, but also whether there exists an auto-regressive relationship between adjacent time points. Running the following analysis for our datasets, we have ascertained that an alphabet size of three is sufficient for us to extract relevant information from all of our datasets.

Selection of Patterns

After the initial clustering step, the second step is to select a subset of patterns or clusters that can be used to represent the system. The underlying hypothesis associated with this method is that while all of the genes have been clustered, only a small subset of them are required to characterize the response of the system. Therefore, the primary question that must be answered is how one characterizes the systemic response. To characterize the response, we propose the creation of a metric called the transcriptional state. Conceptually, the transcriptional state is defined simply as the deviation of the system from its homeostatic baseline. Thus, the motivation behind the selection is to maximize the presence of this deviation.

Given the nature that homeostasis consists of genes with differing expression levels, rather than looking at how the expression level of a given gene or cluster changes over time, we are instead looking at how the distribution of expression levels change over time. Therefore, the difference in the transcriptional state requires a method that quantifies the differences in distribution. For these purposes of this work, the Kolmogorov-Smirnov statistic was used[52]. The Kolmogorov-Smirnov statistic is a simple approach that allows us to quantify the difference between statistical distributions without requiring the use of named distribution. This is an important component because given a population of genes, there are no guarantees that as the system evolves that it will conform to a named distribution[53].

The Kolmogorov-Smirnov statistic however is sensitive to parameter changes between different distributions. Therefore, it is able to distinguish between two Gaussians that have different means or standard deviations. However, for our work, we are more concerned about the how the overall type of distribution changes. Therefore, we have implemented a double normalization operation which z-scores a population of selected genes, first across time, then for each time point. The effect of this operation is shown in **Figure 9**, and allows us to determine for instance how the distribution between high and low expression levels change as a function of time, rather than changes in the underlying parameters of a distribution.

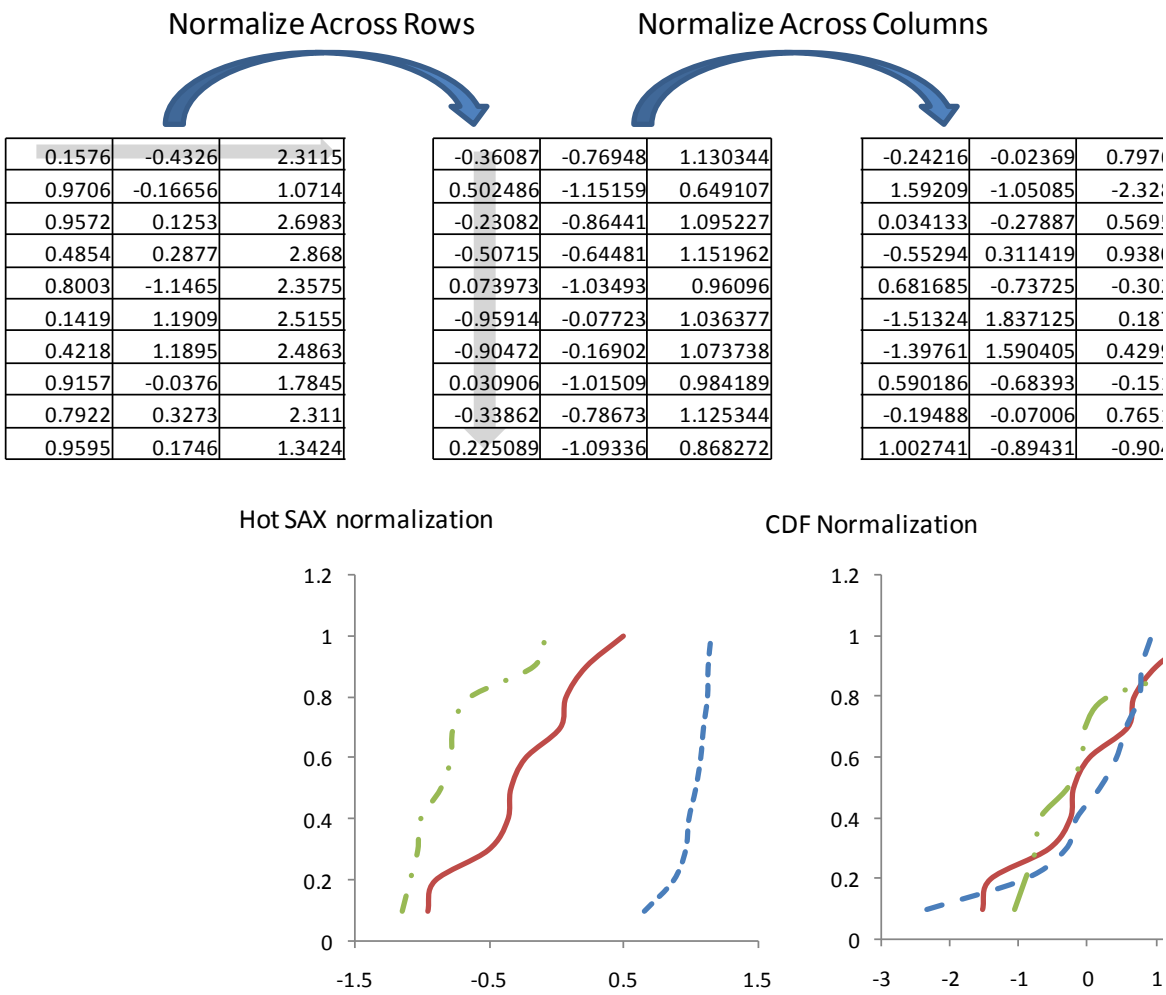


Figure 9:The effect and the justification for the double normalization. What we wish to determine is determine whether two distributions are different in their underlying distributions rather than due to changes in parameters, i.e. Gaussian vs. Exponential rather than two Gaussians with different means or standard deviations

Because the focus of this algorithm is specifically to process time series data, the Kolmogorov-Smirnov statistic has been extended to deal with time series. Therefore, the transcriptional state over time is defined as (2.4).

$$D(t) = \max_{1 \le b \le B} \left| CDF_b(t) - CDF_b(0) \right| \quad \forall t \tag{2.4}$$

Where D(t) is the Kolmogorov-Smirnov statistic is calculated for every time point with respect to the zeroth time point $CDF_b(0)$ which functions as a control. Therefore, there interest is to obtain the subset which gives the greatest deviation, leading to a change in the equation to:

$$\Delta = \max_{t} \left(D(t) \right) \tag{2.5}$$

Where Δ is the maximal difference over the time points.

The selection process attempts to maximize this difference in the KS Statistic. Currently, the selection is conducted in a greedy manner. This algorithm is termed a greedy selection because it does not handle the selection in a globally optimal manner, but instead pre-ranks the different motifs prior to selection. This greatly cuts down upon the number of motif combinations which need to be accounted for, greatly simplifying the overall selection. Thus to obtain a the set of motifs that are representative of the underlying system dynamic, the most highly populated cluster is selected first, and evaluated for its deviation away from the baseline, then the second most highly populated cluster is selected, and the set of both are evaluated for their ability to exhibit a deviation from the baseline. This process is performed until all of the separate motifs have been added. After the transcriptional state has been evaluated for all of the clusters, the subset of clusters which yielded the maximum deviation is selected as the informative subset.

The overall pseudo-code of the algorithm is given below, where S(k) represents the set of motifs being evaluated, k represents the number of clusters added, D(k) represents value of the transcriptional state at a given number of clusters, and Σ represents the informative set.

(i) $k = 0, S(k) = \varnothing, D(k) = -\infty$, max = -∞

(ii) $k = k + 1$

(iii) $h^* = \arg\max N(h)$, N(h) = number of genes with corresponding hash value h

(iv) $G(k) = \{g_i : hash(g_i) = h^*\}$, the subset of genes that hash to h

(v) Evaluate $F(Y_{g_i}(t)); t = 0, \mathrm{K}, T; g_i \in \Sigma$

(vi) Evaluate $\mathrm{D(k)} = \max\left[\max_{1 \le b \le B} \left|\mathrm{CDF_b(t) - CDF_b(0)}\right| \forall \mathrm{t}\right]$

(vii) If D(k) > max

(viii) Max = D(k); F = k;

(ix) Go to (ii) until all peaks have been added

(x) For a = 1 to F

(xi) Select $\Sigma = S(a-1) \cup G(a)$

Results

Plotting the performance of the HOT SAX algorithm upon the chronic and acute datasets, we can establish the optimal parameter for AB. The AB which corresponds to the minimum correlation with the exponential distribution and therefore the furthest from the null distribution is 3 as shown in **Figure 10**. Therefore, the clustering result corresponding to an AB of 3 is used to process the two different datasets. From the observation of the population distribution of the individual clusters for the different datasets, it appears that a significant perturbation has occurred in both dataset. This is encouraging because we see evidence of

significant coordination between the different genes as indicative of the effect of the external drug administration.

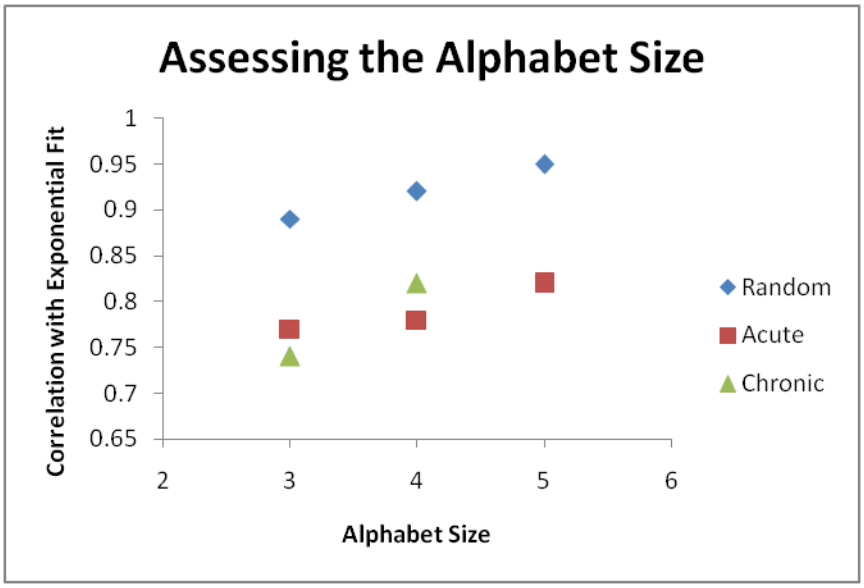


Figure 10: The correlation with the exponential fit of our two datasets compared to a randomly generated dataset. Because an AB of 3 corresponds to the lowest correlation for our two datasets, the value 3 was utilized.

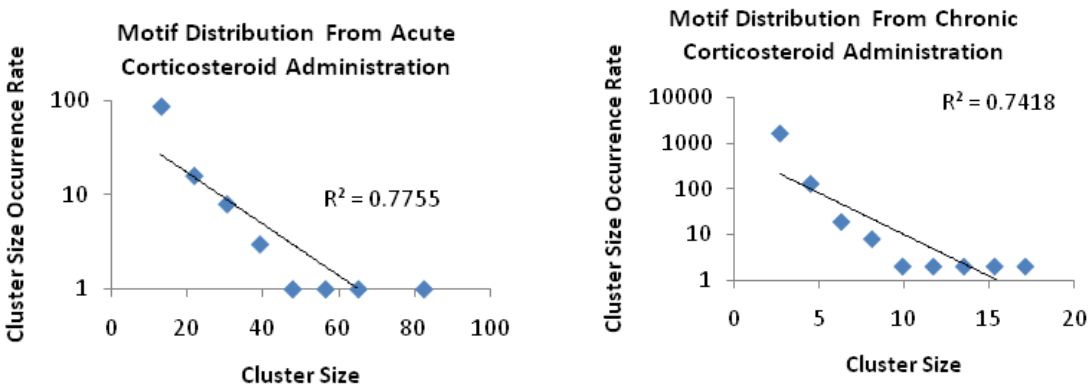


Figure 11: The population distribution associated with the two different datasets after HOT SAX. What is evident is the deviation away from the exponential distribution in both of the datasets showing significant amounts of coordination between the two datasets.

The result of utilizing HOT SAX for clustering can be represented as a histogram. This histogram was generated with a bin-size of one such that each individual hash value is treated separately. This allows for a quick visual inspection to determine which clusters happen to be over-represented. Furthermore, this clustering result is quickly and efficiently generated in O(n) time, and allows us to generate the quickly map over-represented clusters to a set of genes for use in the selection process later on. As we add each cluster into the set of informative genes, we can see that as more clusters are added, the maximum value of the transcriptional state D(t), increases until it reaches a maximum at some intermediate number of clusters. After this point, as more clusters are added, we are essentially adding noise to the system by either adding clusters which show strong similarities to previously added clusters, or through the addition of clusters with only a few genes. Thus, by taking the motifs associated with the maximum value of the transcriptional state, our selection is made **Figure 12**.

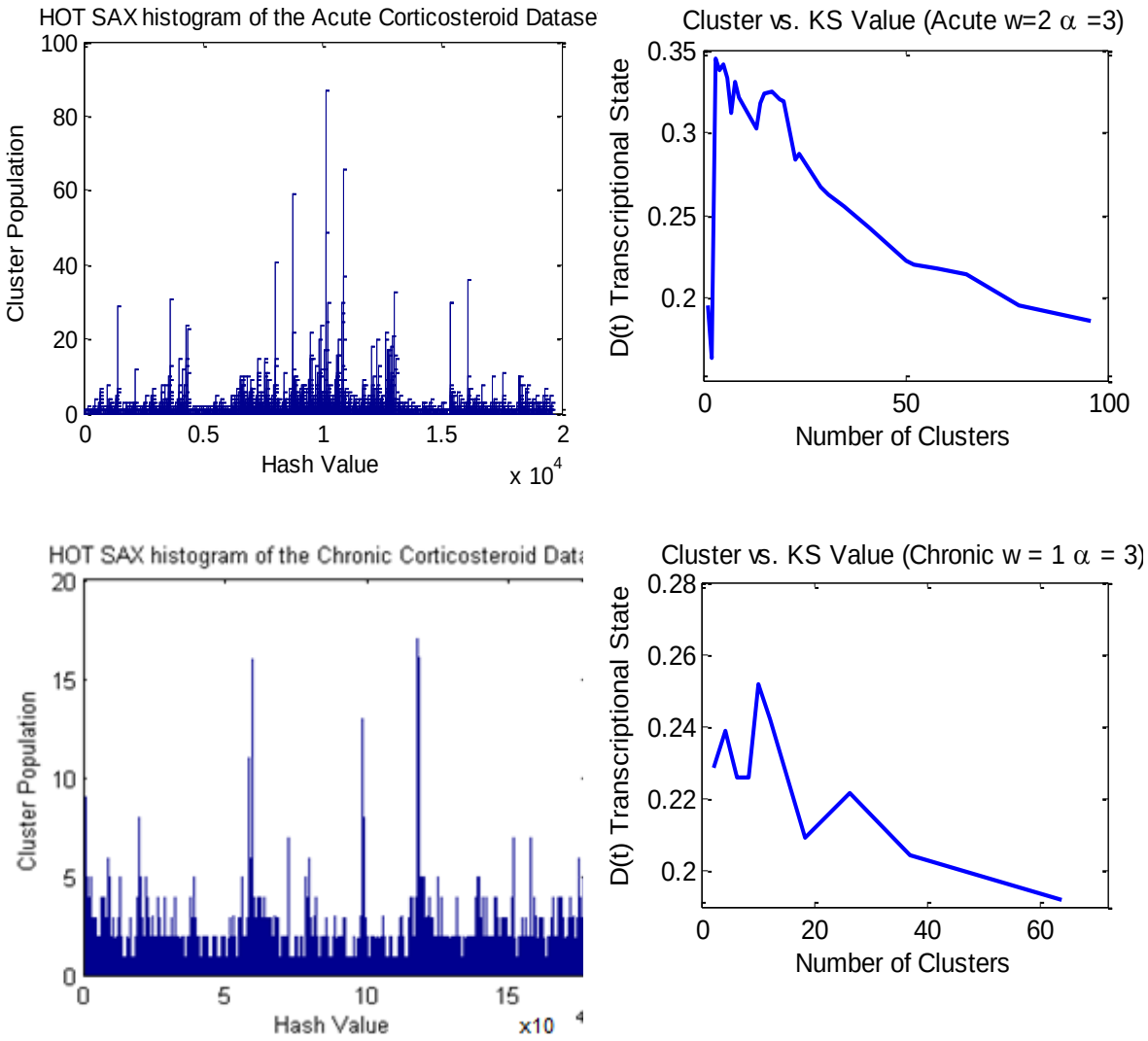


Figure 12: The result of the hash based clustering on the right for both datasets, and the maximum of the transcriptional state over time for various numbers of clusters. By looking at this figure, it is possible to identify the optimal number of clusters with which to represent the system

Rather than looking at the maximum of D(t) as was done in the selection process, there is also some value in looking at the dynamics of D(t) associated with the informative subset. This represents the deviation away from the baseline, and we term this the systemic response. Analyzing the systemic response of the acute corticosteroid case, the dynamics appear to mimic the response of a 2^{nd} order system with respect to an impulse function. Rather than being directly affected by the drug concentration in the system which was modeled as a bi-exponential decay function by Almon et al.[14], what we see is a time lag before the systemic response reaches a maximum, and decay back to baseline after the drug has been cleared from the system. Finally the systemic response of the system to an infusion of corticosteroids appears to follow a two wave response. The primary response appears to mimic the same response observed under the acute case, with a delay before the effect of the drug is maximal and a return to baseline. However, before the system can return to baseline, a second sustained response occurs. One of the questions which arise from this analysis is how precisely the underlying mechanism can give rise to these dissimilar responses. These responses are shown in **Figure 13**.

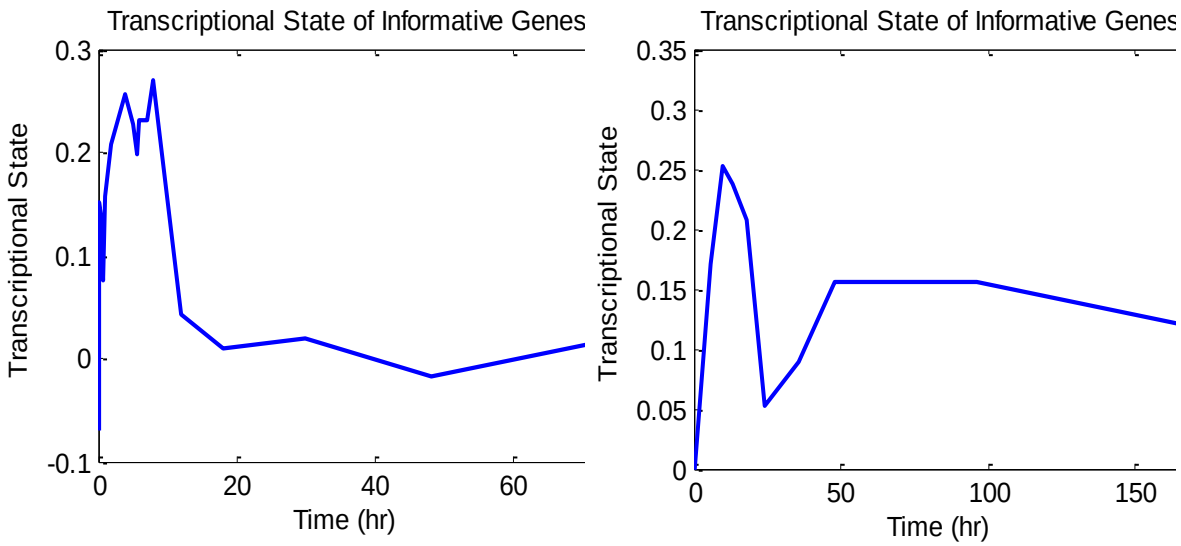


Figure 13: The progression of the transcriptional state of our two datasets. The acute corticosteroid dataset shows a response similar to that of a 2nd order system in response to an impulse stimulus (Right) . The chronic corticosteroid dataset shows a response whose early phase seems similar to that of the acute administration of corticosteroid, but shows a secondary response which is sustained.

As for the specific genes that were selected, the results of the acute selection consisted of 3 clusters which corresponded to 211 genes divided up into three clusters. The responses of these clusters appear to mimic the systemic response as quantified via the transcriptional state. All of them show a 2^{nd} order response in which there is a deviation from the baseline and a return. However, while the two up-regulated clusters appear to have very similar dynamics, they differ in one critical aspect which is time constants associated with each event, with the genes in cluster one appearing to deviate and return at a faster rate than cluster two.

The results of the SLINGSHOTS algorithm upon the dataset corresponding to the chronic infusion of corticosteroids yielded 136 genes divided up into 10 clusters. However, the response of the genes under the chronic administration of corticosteroids appears to comprise up of two different responses. The first response appears to be very similar to the response obtained via the acute dosing in that there is a deviation from baseline, and then a return. However, there exists a secondary response which appears to contain a time lag, before ramping up to a new steady state. The genes which were selected under the two runs are shown in **Figure 14**, and those with known biological functions are tabulated in **Appendix A**. The probe sets which were selected which did not correspond to any known gene or functionality was not included because they are uninformative in deciphering the processes associated with each drug administration.

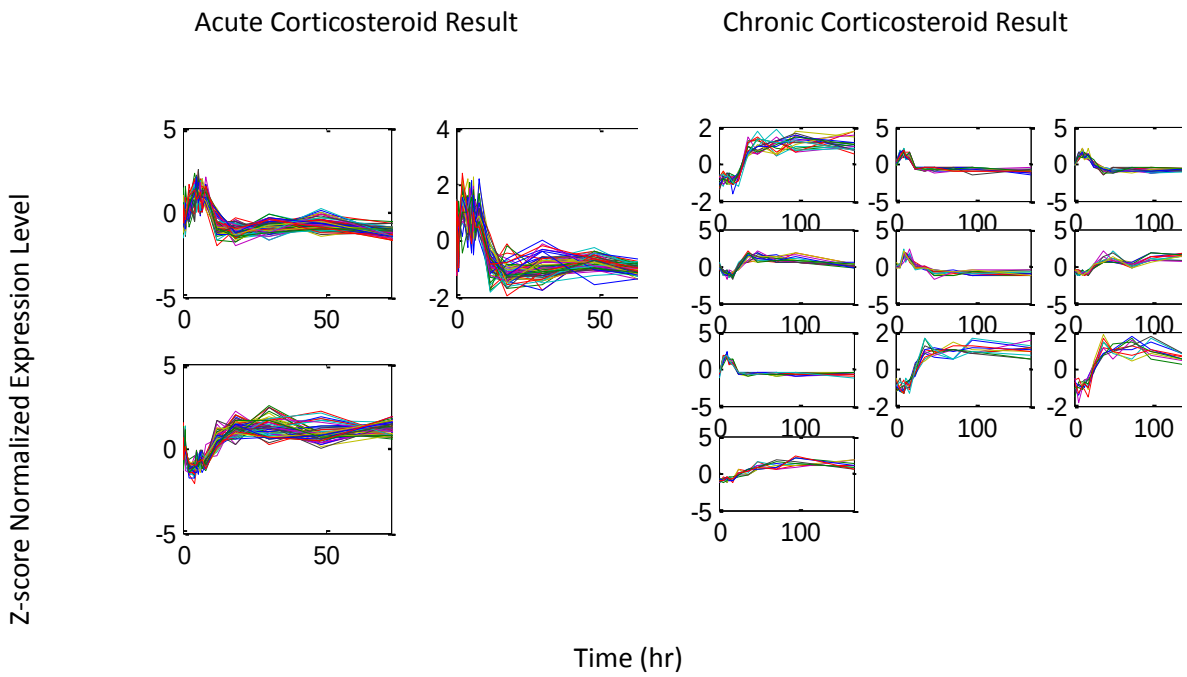


Figure 14: The profiles of all the extracted Genes. Under the case of acute corticosteroid activity (left), we see a deviation away from the baseline followed by a return. In the case of the chronic, administration of corticosteroids (right), we see two distinct profiles, one which returns to baseline and the other one which does not

Discussion

While the response of the curves appear to show the informative nature of the selection, we must definitively establish the fact that the resultant curves are due to our selection strategy, rather than due to chance. To evaluate this, we conducted a random selection of motifs such that the same number of genes were selected as was in the non-random selection. After the selection of these genes in a random manner, we plotted the associated transcriptional state over time. In **Figure 15,** we can see that first there does not appear to be any coherent signal. Secondly, it also appears that the max of the transcriptional state is nowhere as high as when our selection strategy was carried out. Because of this result, we are reasonably sure that the effect which we are observing is not due to the reduction of the number of genes which are being analyzed, but due to some inherently informative gene selection procedure.

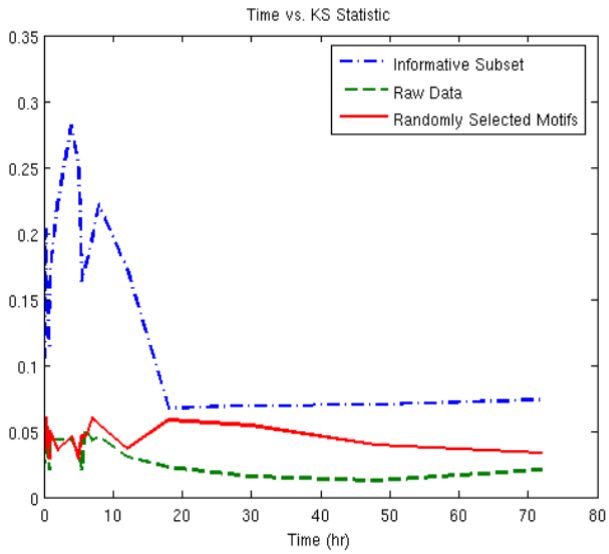


Figure 15: When a random selection of motifs occur, we do not see a large change in the transcriptional state as quantified by the KS Statistic over time, nor a profile which is particularity informative

One of the issues that must be tackled is the question whether the selected genes are actually representative of the intrinsic dynamics which are associated with corticosteroids. In the acute case, the systemic response of the system appears to correspond to the activity of TAT[10], a marker gene that was selected in the preliminary analysis by Almon et al. Furthermore, there does not appear to be the selection of any gene expression profiles which are significantly different from those of the initial selection. Therefore, at least for the results of the acute administration of corticosteroids, it appears that there is good agreement in the dynamics of the system. In the initial analysis of the chronic dataset, Almon et al., focused upon the genes that were selected under the acute case and found that genes which were correlated under the acute dosing did not necessarily have to be co-expressed under the chronic condition[19]. However, from this analysis, it was not clear whether all of the different temporal dynamics have been isolated.

The lack of a gold standard with which to compare our results to is encouraging, because it means that the algorithm will be involved in the synthesis of new information, rather than the validation of previously obtained information. However, in light of this, the validation of the results is more involved. Rather than focusing specifically upon the dynamics associated with the selected genes, or the transcriptional state, we will instead utilize an evaluation of the genes and their associated biological function. By utilizing the ontologies of these respective genes, we are able to link the dynamics of the genes, and various clinical observations associated with these specific genes. A list of all the genes isolated by the SLINGSHOTS algorithm under the acute case and their associated functions are given in **Appendix A.**

Under the acute case of corticosteroid administration, Cluster 1 which exhibited a faster rise and fall, were primarily associated with signaling, whereas the metabolic responses were primarily

found in clusters 2-3 which exhibited a slower dynamic. While the segregation isn't perfect, this result suggests that the metabolic effects associated with corticosteroids lie downstream of the initial activation. What is notable about the results of chronic administration of corticosteroids in **Appendix A** is that a majority of the genes which were related to metabolism where associated with the profiles which showed sustained up-regulation in response to a chronic administration of corticosteroid, whereas the genes normally associated with inflammation were much more likely to be associated with the gene expression profiles which showed an initial up-regulation and then a return to baseline, as predicted via a receptor mediated indirect effect mechanism that showed tolerance. However, the metabolic effects do not appear to show significant tolerance and reach a new steady state with the infusion of corticosteroid administration. This is in agreement with the clinical observations that chronic dosing of corticosteroids appears to have significant prolonged metabolic side effects[54] whereas the overall anti-inflammatory or immuno-suppressive effects appear to be transient[19] such that the prolonged infusion of corticosteroids does little to blunt systemic inflammation and sepsis[55]. Utilizing this information we hypothesize that an infusion of corticosteroids will lose its effectiveness in mediating the inflammatory response, while having a sustained effect upon various metabolic systems. Therefore, the utilization of an infusion of corticosteroids may not be a viable therapeutic strategy if one were attempting to minimize the metabolic side effects.

Coupling the two pieces of information, it appears that the metabolic effects appear to have a significant lag, after the initial immune related responses. From these results we make two primary hypotheses, the first is that many metabolic responses occur as a secondary event, i.e. occurs after some initial response to corticosteroid dosing. Furthermore, this suggests that the previously hypothesized tolerance mechanism associated with corticosteroids may not play a role in limiting the metabolic response. Thus, we hypothesize that rather than a global tolerance

mechanism, there may exist some mechanism which reduces the activity of corticosteroids upon a subset of genes. This piece of information is quite significant, because it suggests that it may be possible to isolate the anti-inflammatory effect of corticosteroids from the metabolic effects of corticosteroids. While at this point, we have not identified a mechanism for doing so, the segregation of the responses hints at the possibility that separate mechanisms may be in play.

Initial Model of Corticosteroid Activity

While different models of corticosteroid activity have been proposed, none of them have been able to replicate the dynamics observed in the results of the SLINGSHOTS gene selection under the chronic case, nor are they able to replicate the observed clinical response to corticosteroids[54]. Though this may point to an issue with the SLINGSHOTS algorithm in general, the fact that these dynamics correlate well with clinical observations suggests that the dynamics which have been selected, are not artifacts of the algorithm and are actually underlying responses of the system which one need to take into account. However, rather than taking a previous model as the starting point for our initial model building, we have elected to start from a simple compartment model of drug activity. This model allows us to explore the question as to whether the response of the system is mediated primarily by the local nuclear concentration of drugs within the system, or whether there exists some significant nonlinearities which must be accounted for.

Previous work in pharmacokinetic and pharmacodynamic (PK/PD) modeling has suggested the applicability of a compartment model with which to model the dynamics of gene expression [56]. The compartment model assumes that the transcriptional activity of a given gene is directly related to the amount of “signal” visible in the nucleus. In this work we will assume that the observed transcriptional dynamics, expressed as D(t), is the manifestation of the activity of an

ensemble of gene responding to an internal mechanism of an external perturbation. The end point of this cascade is the response of the effect which we are interested in. In our specific case, this effector represents the changes in mRNA expression level. The systemic dynamic is expressed as follows:

$$\frac{dX_1}{dt} = I(t) - k_{1,1}X_1 + k_{2,1}X_1$$
$$\frac{dX_j}{dt} = k_{1,j-1}X_{j-1} - k_{2,j}X_j \quad j = 2,\ldots,M$$

(2.6)

The advantage of utilizing this model is that even though it represents a general class of systems, it has specific mechanistic consequences. Our general intuition about the order of the model (number of elements) is drawn from Ockham's razor, and we therefore hypothesize that the model with the smallest number of compartments which can successfully fit the dynamics of the data should be used. For the determination of what should be modeled, in the specific case of our algorithm, it is possible to model either the global dynamic of the system or the dynamics of individual motifs (clusters) that make up the overall response. We can elect to model the global dynamics if we seek to determine how the overall system is responding to the input stimulus or the expression motifs if we wish to determine whether specific differences are present in the different clusters.

We have selected to model the global dynamics of the system, rather than the individual clusters. This was done because of the desire to identify how the system responds to a drug administration, rather than the response of only a single system. Thus, rather than utilizing a single gene as the marker for a drug's activity, we will utilize the transcriptional state as a more

comprehensive marker, one which quantifies in a single vector, the contribution of all the different systems. Thus we will be attempting to minimize the difference in the predicted curves and the transcriptional state obtained from **Figure 13**. Therefore, the objective function which we would attempt to minimize is given in (2.7), where D*(t) represents the prediction from the model, and is represented as the inner most compartment CS' corresponding to the place where the production of mRNA occurs, and I(t) represents the input driving the system. Normally, this term can be thought of as a mathematical surrogate describing how the drug has been administered into the system.

$$\min : \left\| D(t) - D^{*}(t) \right\| \tag{2.7}$$

Thus in the case of the acute administration of corticosteroid, I(t) can be represented via the dirac delta function δ(t), in an infusion of corticosteroids, I(t) is represented as a step function μ(t).

In the case of the acute administration of corticosteroids, the input stimulus is treated as an impulse or the dirac delta function. Fitting the overall dynamic to the global response the results in ***Figure 16*** are obtained. As seen in **Figure 16,** the fit denoted via the dotted lines corresponds qualitatively to the KS statistic vs. time. This corresponds to a generative equation given by (2.8).

$$\frac{\mathrm{dCS}}{\mathrm{dt}} = \mathrm{I(t)} - .1975 * \mathrm{CS} \tag{2.8}$$
$$\frac{\mathrm{dCS'}}{\mathrm{dt}} = .1975 * \mathrm{CS} - .3631 * \mathrm{CS'}$$

Where CS is the amount of corticosteroid in the first compartment and CS' is the amount of activated corticosteroid receptor which is directly affecting mRNA production. The assumption

therefore in this model is that the response of the system is directly related to the local concentration of the drug that is visible to the cytosolic glucocorticosteroid receptor. In this equation, we see again that there exist two compartments within the system which buffer the amount of activated corticosteroid receptor in the cytosol. Because of this buffering effect, the injection of corticosteroid does not have an immediate effect that decays over time, but rather than effect which has lag period before the effect of the drug administration is maximal, and a slow decay as the effects of the drug decay in the system. This was verified in the original experiment, where it was found that the drug was cleared from the circulation after 6 hours. However, in spite of the lack of drug within the circulatory system of the animal, it was found that there was still a significant amount of mRNA activity associated with the system[14].

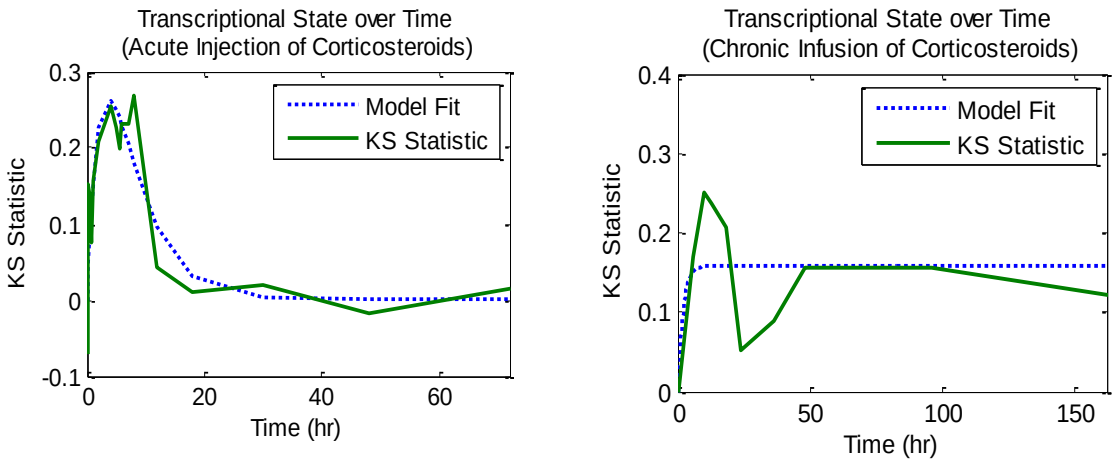


Figure 16: Utilizing a simple two compartment model of drug activity, we are able to replicate the response of the animal model to an acute injection of corticosteroids (left), but not of a chronic infusion. This indicates, that under the chronic administration of corticosteroids, there may be a process other than simple transport which plays a role.

Biologically, this would correspond to the fact that the administration of corticosteroids needs to be transported through two compartments. The first compartment which buffers the activity of corticosteroids is the circulatory system, and the second compartment which buffers the activity of corticosteroid is the cell itself. Therefore, while we have hypothesized that there are two primary compartments which are the circulation and the cell and the activation of the glucocorticosteroid receptor, the transport of this activated transcription factor into the nucleus probably occurs relatively quickly with respect to the transport processes in the other compartments, and therefore has a small effect upon the rest of the system. Thus at this point, it appears that this simple model of corticosteroids with two parameters is able to replicate the dynamics associated with the much more complex fifth-generation model.

A logical progression from the modeling of acute corticosteroid is the determination as to whether the response of the chronic administration of corticosteroids could be predicted from the same type of model used for the acute corticosteroid administration, except that the input has been changed from the Dirac Delta to the step function. What we find is that the model is unable to predict the response associated with the chronic stimulation of corticosteroids **Figure 16**. Furthermore, refitting of the model is also unable to replicate the two-wave effect, with the first set of events begin to subside before a secondary long term response takes over. This suggests that the structure of the model is insufficient in explaining the system's overall response to a dosing of corticosteroids.

This D(t) response corresponds well with what is observed clinically, with the anti-inflammatory effects of corticosteroids mediated by a clear tolerance mechanism[54], as well as the sustained metabolic effects to the organism such as sustained elevated glutamine level and loss of muscle mass[57]. Thus, in a single marker, D(t), we have captured the two disparate temporal effects of

corticosteroids, something which cannot be explained by modeling the response of a single gene. Furthermore, while this curve does an adequate job in describe the clinical observations associated with long term administration of corticosteroids, it cannot be described via the compartment model proposed in (2.6). Thus, the systemic response of an organism to corticosteroids does not appear to be solely mediated by the presence of the drug within the cytosol. However, the inability of the model to either predict or fit the chronic response is not a negative result, because it indicates two important aspects about corticosteroid administration, specifically that due to the nonlinear response of the system, the response of the acute corticosteroid administration cannot be used to predict the chronic response unless *a priori* information such as previously discovered nonlinear mechanisms such as tolerance were implemented[19]. Secondly, the lack of a fit also indicates that the response of the system is mediated by more than the amount of the drug present in the system or in a specific compartment. This may be indicative of more complex secondary signals present in the system which we have not taken into account. From this modeling exercise utilizing the global dynamics for both the chronic and the acute data, we can show that from the gene expression dynamics alone, there is insufficient information to predict the response of the chronic administration of corticosteroid from the acute administration. It is hypothesized that factors such as receptor saturation, as well as secondary signaling factors are not evident during the acute administration of corticosteroids whereas they are clearly affecting the system under chronic administration.

One may be quick to dismiss the model because of its general simplicity. However, what we are able to establish is that there is at least one set of responses within the system that are driven by the concentration of drug in the cell nucleus, i.e. mediated primarily by the transport of the drug. Secondly, the modeling exercise has indicated the fact that despite the fact that one of the responses appear to be consistent between the chronic administration of corticosteroids and the

acute administration of corticosteroids, specifically the responses which deviate and return to baseline, these actually represent two separate types of responses.

Conclusion

The SLINGSHOTS algorithm, when applied to the two different corticosteroid datasets was able to extract temporal gene expression profiles which correspond to the response of the organism to corticosteroids. Therefore, on a superficial level, one may only be interested in the set of genes that has been returned by the algorithm for further exploration. Thus, the large set of genes in the microarray has been significantly reduced to a more manageable number of genes. However, even this set of genes may be too large. However, by examining the profiles associated with these genes as well as the functionality in an aggregate sense, more useful information can be extracted. For instance, the extracted profiles correspond well with the expected physiological response associated with drug administration. In the case of the acute response, we see a clear trend corresponding to a second order model which corresponds to the indirect effect model originally proposed by Jusko et al. In this model, the local drug concentration within the nucleus is the primary driving force that governs mRNA synthesis/repression. Because of the effect of multiple compartments, when the drug has been fully cleared from the circulation, the drug is still present within the cytosol exerting its regulatory influence. This response agrees well with the previously proposed model of corticosteroid activity.

The chronic administration of corticosteroids however does not conform to the previously proposed model of corticosteroid activity. However, while it does not agree with the prior model, it does agree with clinical observations in which there are sustained metabolic effects associated with corticosteroids even though the initial anti-inflammatory/immuno-suppressive effects of corticosteroids are no longer present. Therefore, because of the agreement between

the dynamics of the extracted genes and the underlying clinical observations, we believe that these profiles represent a good starting point from which to create a new model for corticosteroid activity. Utilizing a very simplified compartment model of corticosteroid activity, we concluded that in the case of the chronic infusion of corticosteroids, there is one profile which is directly related to the drug concentration within the cytosol, whereas the secondary profile which was initially attributed to tolerance has another factor involved. At this point, we have identified a starting point for the creation of predictive models because we have identified a basis set from which a proposed model must be able to replicate. Thus, any models which are proposed to model corticosteroid activity must be able to replicate these dynamics. One may be tempted to argue that the results of the gene selection have merely confirmed the clinical observations which have been previously made. However, to create a predictive model of corticosteroid activity, it is important for us to quantify the response of the system in a precise manner, thus necessitating the identification of temporal patterns within our data.

Given the failure of the both the current model of corticosteroid activity as well as our simplified model, the next task is to identify possible mechanisms that can be used to explain such as phenomenon. Given the presence of two divergent families of signals, we propose two possible hypotheses for further exploration. The first question is whether the divergence in responses is caused by the effect of other transcription factors aside from the glucocorticosteroid receptor, and secondly whether a better understanding of the dynamics of the glucocorticosteroid receptor may shed insight as to what the underlying mechanism ought to be.

However, while SLINGSHOTS has shown itself to be successful in the analysis of the two datasets, we feel that improvements can still be made to the algorithm. The most obvious improvement which should be made is in the process of motif selection. Rather than conduct a simple greedy

selection algorithm as was done, one improvement which could be made could be the use of more robust optimization techniques such as simulated annealing or genetic algorithms to conduct the selection. The use of a better selection technique may minimize the presence of multiple motifs which similar profiles, thus allowing us to more systematically assess the number of different patterns within the data. Secondly, because the goal of the algorithm was the selection of significant patterns, it may be useful to also use the selected patterns as template for a more complete and robust selection of overall genes such that a comprehensive set of genes which are responsive in a given biological phenomena are selected.

Equation Chapter (Next) Section 1Selection of Marker Genes

Though the analysis of the microarray dataset yielded a set of genes which response to an administration of corticosteroids, we recognize the fact that many researchers are interested in utilizing high throughput techniques such as microarrays for the selection of biological markers for more targeted experiments later on. Thus, utilizing the entire microarray and the associated transcriptional state may not be a viable option as a biomarker. Because even though the results of the SLINGSHOTS algorithm has reduced the number of possible genes to measure by a large amount, the fact remains that many genes are still in the list. Further complicating the manner is the fact that many of these genes are highly correlated. Therefore, it may be important for experimentalists to identify one specific gene that can be measured through techniques such as RT-PCR[20]. Because one of the goals of our work is the creation of a smaller set of hypotheses which can be tested with further experimentation, it is important that our work be translated into a form that makes further experimentation easier. In any experiment, the most important question is what the assay will be i.e. what to measure and why to measure it. For most traditional works, researchers have hypothesized the existence of a biologically important system and have devised different methods of measuring the activity of such a system. Such markers would then be used either for diagnostic purposes or for the creation of models as was done previously in the case of corticosteroids. However, if one were to select a specific gene for model building, one possible question that arises is, “Given all of the selected genes, why that particular gene?” and specifically whether the gene has been accurately measured such that model building can be performed. Additional complications to the question involve the fact that many of the selected genes appear to have very similar gene expression profiles, so much so that after normalization, it is difficult to tell two genes apart via their dynamics.

Researchers have selected genes which were known *a priori* to play an important role, and the fact that alternative genes exist is immaterial. For instance, tyrosine amino transferase was initially selected as the marker for corticosteroid activity due to its known effect upon protein degradation, one of the effects of corticosteroid treatment[10]. In this approach the high throughput analysis technique reduces the initial set of genes, from which a researcher will apply *a priori* knowledge to select a candidate gene. While all results whether experimental or computational need to be validated by comparing the results with previously obtained results, our overall goal is the creation of computational and experimental techniques that can be used for preliminary examination of a system and thus should not be reliant upon *a priori* information. Thus, because we have already proposed a method for assessing the importance of a given genes, we propose that to distill this larger set into a smaller set of genes amenable to low throughput methods, that we simply select genes that have shown themselves to have been accurately measured.

Our primary hypothesis is that the measurement accuracy of a given gene is dependent upon the following factors: the technical limitations of the microarray platform and the underlying biological variability associated with the process of interest[58]. However, given the fact that most microarray experiments are validated via RT-PCR experiments, we can utilize the converse. Thus, if we were to select a probe based upon how accurately it was measured by the microarray, it has a good chance of being accurately measured under RT-PCR as well provided that the same probe sets are chosen. Furthermore, these probe sets have already been predetermined due to the design of the microarray thus making the selection of probes relatively simple.

We had asserted previously the fact that there is a difference between genes which have been accurately measured and genes that show biological significance. As in the case of the t-test[41], fold change[59], SAM[42], or the ANOVA[40], the question is whether the gene has changed at a statistically significant level. However, in the case of temporal gene expression profiles what is of concern to us is whether the overall dynamic of the system has been accurately measured. Guiding our analysis is the hypothesis that the given quality of a signal increases as the coefficient of variance decreases. In addition, as the number of replicates increases, the confidence is also improved. Thus we propose the creation of new method for quantifying the quality of a given signal which will take both the number of replicates as well as the inter-replicate variance into account.

Method

To satisfy these constraints we propose utilizing a variation of the Leave One Out Cross Validation (LOOCV)[60] technique in which at every time point, either the maximum or the minimum point is removed, and a new ensemble average is calculated. Normally, LOOCV, a specific case of k-fold cross validation is utilized to minimize the degree of over-training or over-fitting of a given classifier or an underlying mathematical model. However, rather than determining whether a given model properly explains the data, we seek to measure the inverse; whether the data reflects the dynamics of some underlying though unknown model. Thus, given the amount of noise present in biological experiments, we seek to verify that sufficient number of replicates were obtained to properly capture the underlying signal rather than noise.

Though there are classes of mathematical models such as b-splines[61] or auto-regressive moving average (ARMA)[62] models which can be used to fit the data, and therefore be used as a basis for the LOOCV analysis, each of them requires some *a priori* knowledge about the

dynamics themselves. For instance when utilizing b-splines, one needs to specify the number of knots or control points to be used by the spline. In the case of ARMA models, the order of the model must be specified *a priori*. In both of these methods, the specification of these parameters will have a significant effect upon how the data is fitted by the model, and therefore a significant effect upon the estimation of how accurate the measured data reflects the underling dynamic. Therefore, we seek a method which is independent of model parameters, and is dependent only upon the confidence interval selected by the researcher.

Leave One Out Cross Validation (LOOCV)

Ideally, we would like to predict whether utilizing an additional replicate for each time point would be change the gene expression profile obtained. While we cannot predict the effect of having an additional replicate, we can simulate the effect by measuring the stability of the signal given n-1 replicates. Thus, treating the ensemble average of a temporal signal as the model, we essentially are evaluating whether taking a subset of the measured data, reflects a similar underlying model. Because the algorithm evaluates a sub-sampled signal utilizing n-1 replicates, this is similar to LOOCV in which one attempts to determine whether a given model can predict the occurrence of a data point which was not utilized in the original training.

Rather than performing the standard LOOCV in which a point is randomly removed from the dataset, we will remove either the minimum or maximum at each time point. Given the small number of replicates normally associated with temporal gene expression datasets, we elected to leave out either the minimum or the maximum point associated with each time point **Figure 17**, to maximize the difference between the different sub-sampled signals. This then allows us to establish a lower bound on the quality of a given signal. Because of this, a signal with length 4 will have 2^4 or 16 possible sub-sampled signals, a signal with length N will have 2^N possible sub-

sampled signals. The primary hypothesis underlying this algorithm is that any arbitrary pair of these sub-sampled signals with n-1 replicates ought to be more similar to each other than randomly generated signals. Thus, for a specific p-value such as $p < .05$, each of the sub-sampled signals need to be more similar to each other than 95% of the pairs of randomly generated signal.

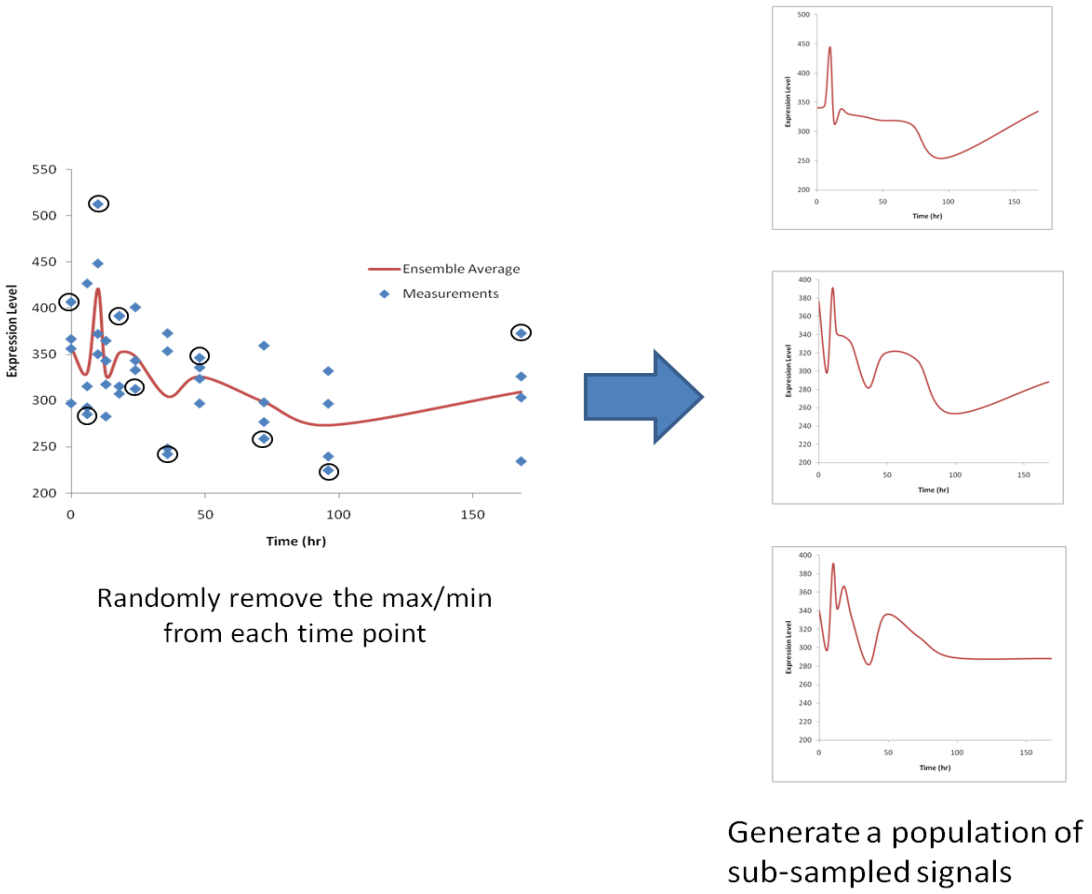


Figure 17: The result of the LOOCV effect upon the average profile of a given signal. Given this random signal, we can see that the removal of one replicate changes the dynamics greatly

Similarity Measure

Given the ability to generate hypothetical gene expression profiles utilizing n-1 replicates, it is then necessary to quantify the difference between these hypothetical signals. To do so, we have utilized Pearson's correlation (3.1) as a method for assessing similarity. Pearson's correlation was selected over other similarity measures because it is scale invariant allowing the comparison of signals of different magnitude. Furthermore, the use of Pearson's correlation is attractive because the R^2 correlation coefficient associated with it can easily be converted into an s-value via (3.2), which can later be converted into a p-value by utilizing the t-distribution[53]. This negates our need to generate a population of random signals to evaluate its quality vs. the null hypothesis.

$$r = \frac{\left[N * \sum_{i=1}^{N} S_1(i) S_2(i) - \sum_{i=1}^{N} S_1(i) * \sum_{i=1}^{N} S_2(i) \right]}{N * \sqrt{\left[\sum_{i=1}^{N} S_1(i) S_1(i) - \sum_{i=1}^{N} S_1(i) * \sum_{i=1}^{N} S_1(i) \right] \left[\sum_{i=1}^{N} S_2(i) S_2(i) - \sum_{i=1}^{N} S_2(i) * \sum_{i=1}^{N} S_2(i) \right]}} \quad (3.1)$$

$$s = \frac{\sqrt{r^2 (N-2)}}{\sqrt{1-r^2}} \quad (3.2)$$

Because of the hypothesis that an arbitrary set of sub-sampled signals ought to be more similar than random genes, this p-value can be used to determine how non-random any fluctuations within the data are. For a given statistical significance threshold, all of the sub-sampled signals will be required to have a correlation coefficient which is more statistically significant than this threshold. Therefore, if the p-value is set at $p < .05$, then all of the sub-sampled signals need to correlate with each other at a level which is more statistically significant than this cutoff.

Assessing the Impact of High Quality Signals

While the motivation behind utilizing this filtering technique was to identify a specific biomarker that was accurately measured, it is difficult to quantify this without performing additional experiments. Due to the computational nature of this dissertation, this was not possible. However, because the data used in the evaluation consists of high throughput gene expression profiles, a surrogate metric can be used. For temporal gene expression profiles, one of the primary hypotheses is that groups of genes with similar temporal progressions of their gene expression profiles will have similar functionalities. Therefore, while the end goal is the selection of a specific gene that can function as a candidate for techniques such as RT-PCR, we will first observe what occurs to a population of gene after they have been filtered under this quality assessment metric. This will allow us to establish the fact that the selection of accurately measured genes can significantly upgrade the confidence in our results.

In our case, we have elected to use the clustering package cluto[35], with the default parameters as a representative clustering approach. This is identical to the analysis approach that was utilized when we were assessing the inherent qualities of the dataset themselves. However, rather than showing the effects of having a good dataset, we will show that it is possible to upgrade the informative nature of each dataset through a filtering technique. Therefore, with an increase in quality of clustering due to better signals, it should be possible to see an associated improvement in the enrichment[29, 63]. Gene Ontology enrichment is conducted by utilizing the hypergeometric distribution as given in (3.3). The ontologies themselves are obtained from the Affymetrix Annotations provided with each individual microarray. This hypergeometric distribution essentially calculates the probability that a subset of genes has been selected from an overall population. To evaluate the overall quality of a given enrichment, the metric will be the percentage of identified ontologies which have been selected as enriched. It is hypothesized

that if the clustering is more reliable, then there should be a lower number of ontologies which had been spuriously included due to ambiguities within the signals.

$$P = 1 - \sum_{k=1}^{n} \frac{\binom{k}{i}\binom{m-k}{N-i}}{\binom{m}{N}};$$

n = number of times the ontology appears in a given cluster
i = number of genes in a given cluster (3.3)
N = total number of genes
m = number of times the ontology appears in the dataset

Given that the initial claim of the manuscript is that it is important to select for genes which show not only significant differential expression, but also genes which show accurately measured expression profiles, thus we have elected to compare the performance of the proposed LOOCV algorithm vs. a standard method for selecting genes based upon differential expression ANOVA[40].

One of the difficulties with this assessment is that the evaluation of gene enrichment is dependent upon the number of clusters with the data is partitioned into. Determining the number of clusters itself is an open area of research, and thus it is difficult to determine the proper number of clusters present within the data. Therefore, instead of focusing upon the number of clusters present in the data, the evaluation will be conducted over a continuum of different cluster numbers. It is hypothesized that if the filtering has been successful, then the percentage of significant ontologies will be greater for any given cluster number.

Results

For all of the datasets, the p-value cutoff was selected at $p < .05$ for both the ANOVA as well as the LOOCV Quality Assessment. While it is arguable as to whether such a threshold is

appropriate given the number of genes present within the dataset[64], what we seek to show is that for a given threshold that filtering genes based upon the accuracy in which their dynamics has been captured has a greater impact upon clustering than just selecting the genes based upon their activity.

For all of the datasets, the selection of genes based upon the quality of their dynamic expression profiles showed a consistent trend in that genes which had been filtered based upon the accuracy of their dynamic response show a greater percentage of enriched ontologies as compared to the genes which passed the ANOVA filter**.** This increase in the percentage of significant ontologies is due to a decrease in the total number of ontologies associated with the selected genes, and not due to an increase in the number of significant ontologies.

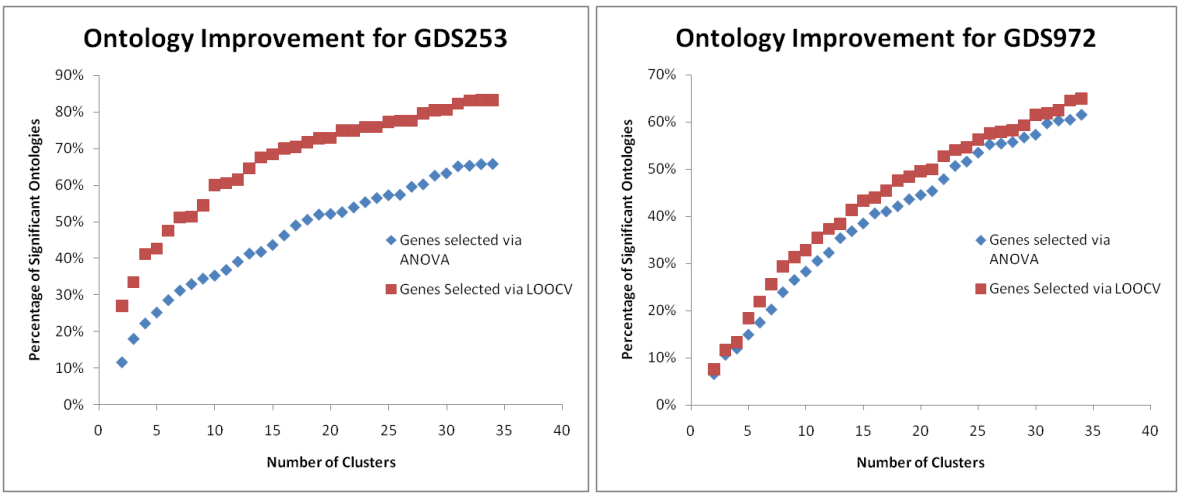


Figure 18: Fraction of identified ontologies which were enriched at a statistically significant level. In both our datasets, there is an improvement in the percentage of significant ontologies when utilizing selection for high quality signals.

For the GDS972 chronic corticosteroid dataset, we see the smallest amount of improvement between filtering the dataset utilizing an ANOVA vs. the proposed LOOCV filtering algorithm. This was predicted *a priori* because of this dataset consists of more replicates as well as the fact that the RAE230A microarray itself has a higher signal to noise ratio than the older RG-U34A arrays[65]. Therefore, because of the higher inherent quality of this dataset, many of the genes which show significant changes in gene expression profiles were also accurately measured. This is borne out by the fact that the intersection between the two sets is quite high with 3038 of the 4361 genes selected via ANOVA also being present in the filtered set. Thus, for all of the genes in which a statistically significant change did happen, it was also quite likely that the gene had also been accurately measured.

The GDS253 dataset was hypothesized to show the greatest improvement from the selection of accurately measured genes. However, aside from the initial prediction as to the ontology improvement, this dataset was surprising because of the fact that 317 out of the 438 genes selected via ANOVA were also found in the LOOCV filtered set. Thus, the removal of a similar percentage of genes had a much larger effect upon the ontology enrichment of the GDS253 dataset. We hypothesize that the effect may be due to the fact that the genes selected by ANOVA which were not presented in the set obtained via LOOCV quality assessment introduced many genes that could not be accurately grouped in any of the clusters, and therefore introduced a large set of ontologies which were not similar to the ontologies of the other genes within a given cluster.

The selected genes for each identified cluster form the SLINGSHOTS algorithm has been provided in **Appendix C**, as well as the average profiles and each of the replicates. As seen in the table, the replicates of all the genes appear to be tightly clustered, specifically that the intra-

replicate variance is much lower than the inter-replicate variation. Thus, the temporal variation of the signal as measured by the mean of the replicates represents an accurate reconstruction of the signal and the response of the underlying cluster.

One of the encouraging results of the SLINGSHOTS algorithm as quantified by this metric is the fact that most of the genes selected by the SLINGSHOTS algorithm appear to be well measured with a significance level of $p < .001$. Therefore, at least we can be reasonably sure that the genes which we have selected as being biologically significant also show significant measureable changes in expression.

Discussion

Ideally, one would have liked to validate that the genes that were selected as being accurately measured could be accurately measured through other techniques such as RT-PCR. However, because the focus of this work has been computational rather than experimental, this validation was not performed. In light of this, we had proposed an alternative method for assessing whether these signals were indeed of high quality, specifically that they increase the overall confidence one has after performing a clustering operation. We have definitively shown that this is the case by showing that genes selected with the quality of their temporal expression profiles in mind show greater gene enrichment than only those that show significant differential expression. In light of this result, it is possible to utilize this signal quality filter to be used for the initial gene selection step rather than the SLINGSHOTS algorithm. This is further reinforced by the fact that the SLINGSHOTS algorithm appears to select for genes that have been accurately measured as well. Therefore, there is some interaction between the two selection operations. However, the use of the SLINGSHOTS algorithm is still encouraged because of its distinct lack of ambiguity when it comes to the selection of parameters, as well as the fact that multiple pieces

of information can be isolated such as the informative nature of the system and the systemic response of the system.

It is important to note that while the set of informative genes selected via the SLINGSHOTS algorithm has been reduced further to a handful of genes, the purpose of these genes is to provide markers for further experimentation. Therefore, if one were to run an experiment at different dosing levels for model validation, one can use low throughput methods to measure the response of the system rather than having to run microarray data. However, while these genes can serve as adequate markers for activity, at this point we do not ascribe particular biological significance to them. It is our hypothesis that in complex phenomenon such as the response of the system to corticosteroids, the ability to measure such a gene accurately is not a good determinant of biological activity, because each gene may play a small role in a much bigger mechanism or due to the fact that we cannot ascribe cause or effect at this level of analysis.

One of the more notable observations which we were able to make was the marked improvement of the newer arrays with respect to the older arrays. Utilizing the newer arrays, we were able to obtain a significantly greater fraction of genes whose temporal responses have been accurately measured. Secondly, we observe a smaller improvement in the gene ontology enrichment after filtering for these accurately measured genes. While this result may not be significant, it does however show that the results of our algorithm do agree with our general intuition that with newer generation of microarrays, the results do in fact improve due to increased signal fidelity. Due to this, we hypothesize that as technology improves, the number of genes that will show differential expression for a given p-value will increase as the SNR increases due to lower spread between the replicates. Thus, while it is necessary to show that a

biologically important gene has been accurately measured, not all accurately measured genes may be biologically significant.

However, while the results of our observation conform to our intuition that newer arrays have better signal to noise qualities, there are complications which we have not been able to fully address. Due to uneven temporal sampling, one significant issue has arisen, specifically how to deal with the samples which encompass a shorter time duration vs. samples that represent the response over a longer duration of time. For instance in the case of the GDS253 dataset, the sampling rate ranges from 15 minutes to 24 hours. Thus while, the majority of the signal in terms of duration of time may have been well captured, the overall correlation coefficient may be low given the high variability in the early time points.

The primary reason for this problem is the fact that the algorithm essentially treats the data as a vector of values without time dependence. Essentially the data points themselves are all given equal weight whether they take place during a short period of time, or whether the data point encompasses a greater period of time. Thus the correlation coefficient or clustering analysis may not also agree with one's judgment utilizing visual inspection of the data. However, while the results of the algorithm may not agree with one's intuition when visually assessing the data, the fact that researchers have selected such an uneven sampling strategy means that the dynamics early may play just as important role as the later dynamics despite their transient effect. Therefore, while there exists algorithms that will normalize the data based upon the time duration via techniques such as interpolation or curve fitting[66], they may miss or minimize the fact that earlier time points may in fact be more important biologically.

Conclusions/Future Work

At this point we have obtained the identity of various genes which can be used to quantify the response of the system for future experimental work, as well as establish the fact that the results of the SLINGSHOTS algorithm as well as being biologically relevant, also show high quality temporal expression profiles. Thus, we can be confident that such results can be replicated under future work. The genes selected are surrogates for the dynamic activity of our system. However, it must be noted that at this point, we make to claims as to the specific biological importance of a specific selected gene. Because the genes were selected based upon technical issues relating to the accuracy of measure, the selection step is not one that determines the biological significance of a given gene. This limitation is one which we think carries over to other selection methods which rely upon differential expression as stated in the previous chapter.

However, one limitation of the work as presented is that experimental validation has not been conducted. At this point, we have not isolated the primer probes associated with the identified genes and run RT-PCR to verify that these selected genes can be measured accurately with alternative methods. Thus future work, should involve utilizing RNA samples obtained from the previously run experiments and validate that the dynamics associated with the microarrays are reflected in the RT-PCR. If it can be shown that genes selected under a certain condition will illustrate similar dynamics under mRNA microarrays vs. RT-PCR arrays, it would obviate the need to conduct a separate wet experiment to validate the results of an mRNA experiment, provided that they satisfy this underlying metric. This would greatly increase the confidence in the results reported via mRNA experiments and their corresponding analysis.

Equation Chapter (Next) Section 1Identification of Possible Alternative Regulatory Transcription Factors

After having obtained a set of genes, and their associated gene expression profiles, we would like to determine just how these genes are regulated. While it has already been established that the glucocorticosteroid receptor plays a key role in triggering the response of the organism to an administration of a corticosteroid, it is unclear whether the action of the glucocorticosteroid receptor may trigger the activation of a secondary transcription factor, which then go on to regulate other genes within the system or whether an alternative transcription factor has a role in regulation. Our inability to model the response of the liver to different methods of administering corticosteroids may be due to the existence of an alternative transcription factor which we have not accounted for. Thus, of interest to us is whether there exists another transcription factor which can be hypothesized to act as an alternative regulator of the genes which we have identified as sensitive to corticosteroid activity.

The binding of these transcription factors has been determined to be sequence specific through various binding experiments[67]. Previous work by Wasserman et al., have shown that this fact can be used to predict the existence of regulatory motifs within the DNA sequence. However, given the relatively short lengths of these recognition sites raging from 6-14 bases[68, 69] as well as the degeneracy possible with each given transcription factor binding site, the probability of a random hit is quite high. More problematic in this evaluation is that the transcription factors can be shown to bind *in vitro* even if they show no *in vivo* activity. This suggests that there exist other conformational factors that regulate whether a given sequence in the DNA is available for binding.

Most researchers have tackled the problem of false positives via the method of phylogenetic footprinting[70-79]. The core assumption in phylogenetic footprinting is that significant control mechanisms in an organism are evolutionarily conserved. Therefore, by utilizing the genomes of multiple related organisms, one should be able to identify conserved regulatory regions within the DNA. The primary benefit of this technique is that it limits the search space for which possible transcription factors binding sites can be found. This technique is exemplified by tools such as CONSITE[80], and FOOTER[73], which look for sequence homologies between two different species. CONSITE represents the basic phylogenetic analysis technique presented by Wasserman et al.,[68] in which only sequences which show high homology between two species such as Rat and Human would be analyzed via Position Weight Matrices (PWM) in order to determine which transcription factors binding sites are present. The primary difference between these and other tools concerns the different ways in which homologous sequences are identified.

In predicting transcription factor binding we explore the notion that “Co-expression implies co-regulation”[81]. With multiple genes requiring similar transcription factor binding interactions, there exists a basis for eliminating false positives. This method allows for the selection of transcription factors binding sites that are active under a given experimental paradigm, thereby allowing us to indirectly incorporate the effects of chromosome and recognition site presentation upon transcription factor binding prediction. Rather than having to rationalize that a few transcription factors binding sites are over-represented in a cluster of genes, one can show that a few transcription factors are active in the cluster of genes that have been grouped together. Although the method focuses on predicting experiment-specific transcription factor binding sites, it is possible that if such a methodology were used in an iterative process where different experiments were analyzed, one could obtain a comprehensive set of transcription

factors binding sites which regulate the various dynamic responses shown by biological systems under a variety of conditions hence building a more comprehensive model of transcriptional regulation. Thus, the general hypothesis is that in a set of co-expressed genes, one can identify factors which co-regulate the genes by identifying the prevalence of a given transcription factor.

Methods

The identification of possible transcription factor binding sites is broken down into two steps: (i) the identification of the promoter region, (ii) the identification of putative transcription factor binding sites. CORG[82] was used for the identification of promoter regions as well the identification of relevant transcription factor binding sites. CORG was selected primarily for its ability to extract the 5' upstream region up to the next gene rather than to a set number of upstream base pairs. This was important to us due to the nebulous concept of how far upstream a promoter region lies. It has been shown that the GRE (Glucocorticosteroid Response Element) could be found thousands of base pairs upstream of the start codon[79]. Other such as TRED[83] on the other hand require as a parameter the number of upstream base pairs to consider. Additionally by using CORG, one is able to utilize its built in facilities to both extract homologous sequences as well as transcription factor binding sites.

One complication which needed to be addressed was the fact that CORG returned homologous sequences between two species and is unable to return just the entire promoter region for a single species. In order to compensate for this drawback, the evaluation was conducted in the following manner. To evaluate the difference between phylogenetic footprinting and our proposed approach of looking at the promoter regions of a set of clustered genes in aggregate, a CORG search was conducted upon human/rat and mouse/rat. The human/rat case is the baseline example of phylogenetic footprinting in which ideally there will be a small set of

regulators which give rise to the similar responses to corticosteroids in humans and rats. The mouse/rat case was used to give a proxy for the context specific case in which the analysis is performed only on the rat promoter region and to determine the transcription factors which are present in all of the genes in the cluster. The rationale for running this case is that the rat/mouse promoter regions have about an 85% conservation rate among homologous sequences[84]. Given this high level of conservation between the two different species as well as the fact that CORG keeps sequences that show a homology of greater than 70% over 100 base pairs[85], it provides a reasonable facsimile for the rat promoter region.

Data Analysis

The primary metric which to be analyzed is the number of times a transcription factor binding site is found in the 5' region of genes that comprise up of a highly correlated cluster. This is necessary in order to determine whether or not there are any transcription factor binding sites which were present in a sufficient percentage of genes where it would be a reasonable candidate for the co-regulation of the genes within the cluster. Secondly, once the metric is quantified, it may be possible to ascertain the overall distribution of transcription factors throughout the cluster of genes, allowing one to determine whether or not the highly conserved transcription factor was present due to a statistically significant event, or whether it was highly conserved due to chance.

The process of finding a hit for a specific sequence in the promoter region can be modeled by an exponential distribution whose PDF is given in **Figure 19**. In **Figure 19**, a random set of genes was selected and the prevalence of a given transcription factor in the upstream promoter region was determined. From this distribution, it appears that the initial assumption that one can model transcription factor occurrence rate on a cluster of gene as an exponential distribution.

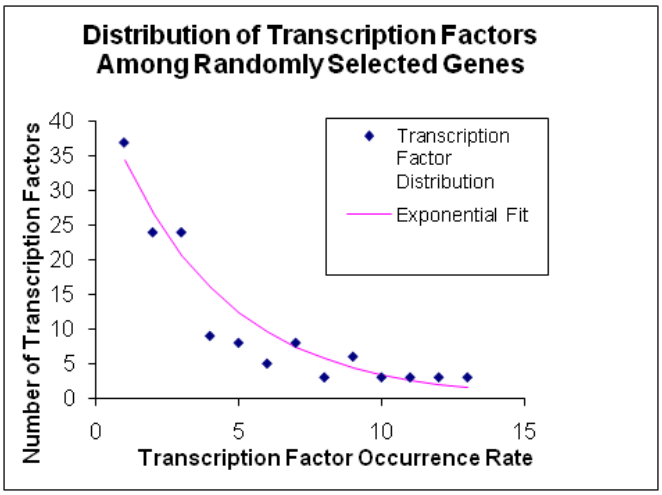


Figure 19: The distribution of transcription factors among randomly selected genes should have an exponential distribution, which is shown above

$$pdf(x) = \frac{1}{\lambda} e^{-\frac{1}{\lambda}x} \quad (4.1)$$

To obtain the parameters for the PDF, the mean number of times a transcription factor binding site is present amongst the genes in a cluster as well as the standard deviation this distribution is calculated. Given the slight discrepancy between the two values, the average of the mean and the standard deviation is used as the parameter with which to model the distributions. The exponential distribution will then allow us to obtain the probability that a single transcription factor will be conserved over x% of the time. This probability will be used below to calculate the expected number of highly conserved transcription factors.

Results

The previous analysis of the temporal gene expression profiles for the acute corticosteroid dataset yielded 3 clusters with 211 genes. This will function as the starting point for our analysis to determine the ability of this algorithm to extract a meaningful set of transcription factors. Previous data that has been presented suggests that for genes to have a greater than baseline chance of having transcription factors in common, the correlation coefficient should be greater than 0.75[81]. Our clusters show an average correlation coefficient of 0.85, comfortably over the limit. This average correlation coefficient allows us to establish a reasonable expectation that one ought to find a set of transcription factors which co-regulate the genes in each cluster. However, when running the analysis upon our three most populated clusters, we obtain the results shown in **Figure 20.** This is a relatively disappointing result to say the least because what is seen is the fact that there does not appear to be significant over-representation of a set of transcription factors associated with each cluster. While some of the transcription factors have shown themselves to be prevalent in many of the genes which make up a specific cluster, their

occurrence is something which can be explained by chance, rather than due to some underlying mechanism.

While it is understandable that this behavior is observed when taking a random set of genes, in which one does not expect co-regulation, we had not expected this behavior in a set of co-expressed genes. Given the hypothesis that co-expression implies co-regulation, we had hypothesized that a much larger fraction of transcription factors ought to have appeared in a set of co-expressed genes. However, what we see is that this is not the case, but rather performance which is qualitatively similar to that of the random grouping of genes.

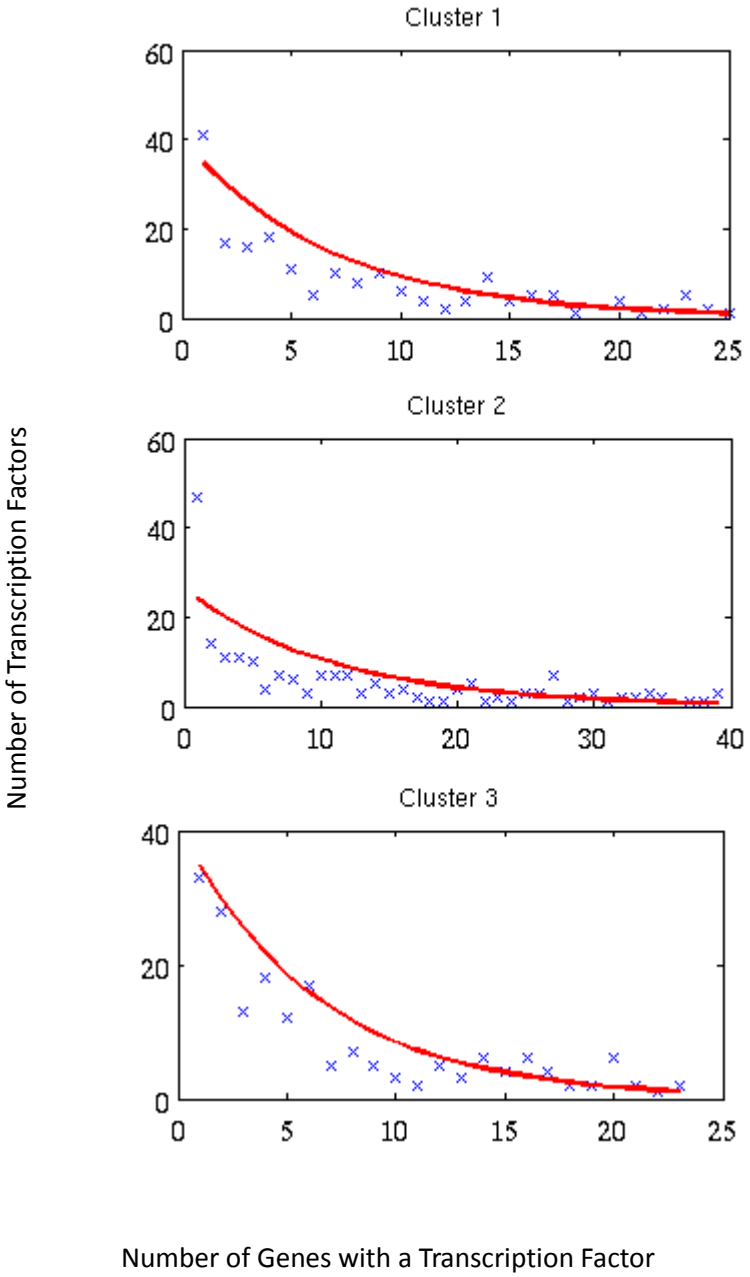


Figure 20: The distribution which shows the prevalence of a transcription factors in a population of co-expressed genes

One possible reason for this may be due to the fact that at this point phylogenetic footprinting has not been applied. It could be possible that the promoter sequences contained too much non-regulatory regions, and that the true response of the system is buried under a lot of extraneous hits. Therefore, we are examining transcription factor hits over what are essentially large stretches of random sequences. Therefore, the exponential distribution would be evidence of this phenomenon. Therefore, the next step in our evaluation is to determine whether the observed dynamic significantly changes if we focus our analysis only upon the set of sequences that are hypothesized to be regulatory sequences. However, after conducting phylogenetic footprinting between humans and rats, we see that the probability of a transcription factor binding to a set of genes can still be accurately modeled by the exponential distribution **Figure 21**. Again, while there may be some transcription factors which are present over most of the genes in a given sequence, their occurrence can be predicted due to random chance. However, one interesting observation is the decrease in the number of hits of a given transcription factor after phylogenetic footprinting. Over all, the number of genes which are regulated by a given transcription factor appears to decrease proportionally to the length of the sequence being analyzed. Therefore, if phylogenetic footprinting reduces the number of bases by an order of two, there is a corresponding two-fold loss in the probability that a given transcription factor will be found to bind to that gene. This suggests that looking at a set of genes which are co-expressed, the use of transcription factor prediction as well as phylogenetic footprinting does very little to bias the system towards a specific set of transcription factors that co-regulate a given population of genes.

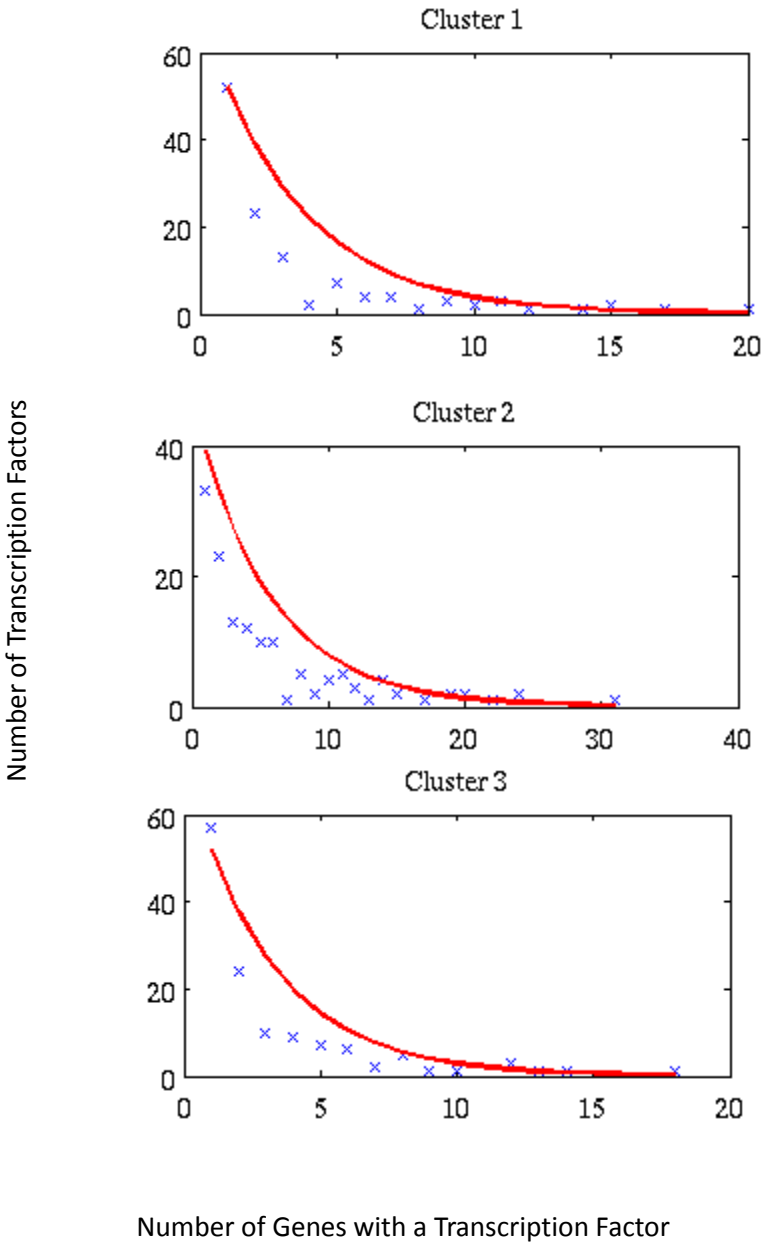


Figure 21: The distribution which shows the prevalence of a transcription factors in a population of co-expressed genes. Even though phylogenetic footprinting has been carried out, we see the same exponential distribution

Discussion

The goal of this analysis was to determine whether by looking at a cluster of co-expressed genes, it would be possible to find a set of co-regulators which would be sufficient in determining possible candidates that would justify the difference in the expression profiles. However, while there are some differences in terms of the over-representation of a given transcription factor in each of our clusters, it is difficult to justify qualitatively their overall importance.

The main point of phylogenetic analysis has been the reduction of false positives in transcription factor binding predictions. We believe that by performing phylogenetic analysis between human and rat as well as utilizing mouse and rat to extract a homologue for the rat promoter region, a bias for true regulatory regions would be introduced, and thus the hypothesis that co-regulation implied co-expression could be verified. However, it has been shown that phylogenetic footprinting does not introduce any sort of bias into the system. We had expected that while there were numerous false positives generated via standard transcription factor binding site prediction that transcription factor binding sites were more prevalent in “true” regulatory regions that were conserved through evolution than over the baseline rate. However, we did not find a greater affinity for transcription factor binding sites to be localized to regions of evolutionary conservation than over that of non-evolutionary conserved segments of the 5’ region. This leads to the hypothesis that the primary driving force in the number of times a given transcription factor occurs within a gene cluster is driven by the length of the promoter region analyzed and the specificity of a given position weight matrix.

However, this disappointing result may be related to the field of transcriptional network analysis. It has been widely noted that maps of transcriptional interactions appear to have a scale-free topography in which the distribution of links between different genes follows an exponential

distribution[86-88]. Additionally, it has also been observed that despite the apparent scale free nature of the network, biological transcription networks illustrate a higher degree of robustness than could be normally explained via a scale free network[89]. Specifically that the removal of a large number of hubs are not lethal to an organism. It has been shown that in yeast, the removal of 28 out of 33 highly connected hubs did not lead to the death of the given yeast cells[89] with little correlation between the connectivity of a node and its importance to viability. Thus, perhaps the complex interplay of different factors may be confounding this relatively simple analysis.

One of the more interesting and notable results was the fact that the glucocorticosteroid responsive element was not found to be over-represented within this set of genes that were hypothesized to be responsive to corticosteroids. This observation has three specific possibilities. The first possibility suggests that perhaps, the genes that are well correlated with corticosteroid responsive genes that were previously identified are responding to a second currently unidentified transcription factor that is sensitive to corticosteroid administration. Secondly, it is possible that these genes are responsive to other regulatory factors which were in turn are directly regulated by corticosteroids. Finally, it possible that the consensus sequence associated with corticosteroids has been inaccurately determined. At this point, with the level of analysis which we have performed, we cannot rule out any of the cases. Thus, further analysis is needed in order to determine which possibility is most likely.

Conclusion

While the results which were obtained were disappointing, we have established a possible link between the specificity of a given transcription factor and the scale free nature of biological networks. Specifically, the promiscuity of a given transcription factor directly relates to its degree

of connectivity within a network. Thus, we hypothesize that highly connected transcription factors will show the greatest degree of ambiguity within its position weight matrix.

Secondly, the result of this evaluation suggests that the sequences which are conserved via phylogenetic footprinting do not show a significant bias for the binding of transcription factors. Thus, the probability of a transcription factor being a match for a given sequence does not change depending upon whether a genomic sequence is phylogenetically conserved or not. Thus, these regions are not biased towards the binding of regulatory proteins if one looks only at sequence data. Thus, if transcription factors are more likely to bind to these sequences, there needs to be other factors at play that govern transcription factor binding aside from the underlying base pair sequence.

The failure of this relatively simple analysis suggests that more outside information needs to be incorporated. Thus rather than focus upon a cluster assignment and sequence analysis may be too simplistic. However, in light of this limitation, it is still important to note that we still have a wealth of other information which has not been incorporated such as the expression profiles of the genes themselves. Therefore, while this method has not been successful in obtaining desired results, it at least points to future avenues which can be explored.

Equation Chapter (Next) Section 1 miSARN for the Identification of Regulatory Networks

One of the results of the previous promoter sequence analysis involved the fact that in many of our extracted genes, there did not appear to be an over-abundance of the glucocorticosteroid responsive element. This led to three different possibilities which needed to be resolved. These included alternative corticosteroid responsive transcription factors, multiple signaling levels, or improper identification of the consensus sequence. Aside from the ambiguities associated with transcription factor prediction we also have the issue of how to properly exploit the information that is obtained. Even if it were possible to isolate without any degree of ambiguity as to which transcription factors regulated the genes in question, we still do not have any idea as to how these transcription factors are able to dynamically affect the levels of mRNA gene expression. Recently, methods combining TF-gene connectivity data and gene expression measurements have emerged in order to quantify these regulatory interactions. Prominent examples are the decomposition-based methods which combine ChiP and microarray data and inversion of regression techniques to estimate TFAs [90-93]. Singular Value Decomposition and regression methods were combined [94] in order to reverse engineer regulatory networks, whereas in [95] promoter elements were linearly combined to quantify the contribution of the promoter architecture on a gene's expression. Network Component Analysis (NCA) [96-100] was introduced as an alternative for quantifying the strength of the regulatory interactions and for elucidating true TFAs, [101] explore a similar linear superposition of expression profiles and TFA combined appropriately using binding affinities in lieu of stoichiometric coefficients and a Bayesian error analysis of an, effectively, linear method was presented in [101]. We will be extending the concepts proposed by NCA in order to identify both the most likely regulators of a given gene, as well as the dynamic interactions of these interactions.

However, while our experimental system is based upon corticosteroids, we will first be evaluating our system upon *E Coli.* This was done to reduce the amount of complexity in our evaluation. We first wanted to determine whether it is possible to obtain the transcriptional dynamics of the system before evaluating how the system is able to deal with ambiguous transcription factor predictions as was done in the previous step. In a nutshell, modeling transcriptional networks will enable us to gain an important insight into the principles that govern the regulation of cellular behavior and gene expression. In this study we model a regulatory network as a mixed integer linear programming (MILP) in which we can incorporate biological knowledge in terms of equality/inequality constraints. Subsequently, our mixed – integer based formulation (MILP) is so flexible that also allows us to generate alternative network structures that account for the same root mean square error (RMSE) or reconstruction elucidating the underlying regulatory rules that govern transcriptional regulation.

In the following section we present our optimization-based formulation followed by its validation with real experimental data for the well-studied organism *E coli*. We chose *E coli* due to the available information from RegulonDB database [102] which characterizes the role of a transcription factor as an activator or repressor. Thus, in the section of implementation we present alternative ways of generating multiple solutions with biological impact. Our algorithm was formulated in GAMS modeling language (General Algebraic Modeling System) [103]. In the section of discussion it is interestingly annotated how such mathematical formulations can shed light on complex biological phenomena such as gene regulation.

Modeling

Network model

Modeling gene regulation as a linear model we assume a quasi steady-state for mRNA synthesis and degradation [101] where transcription initiation can be described by a set of reversible reactions that all reach equilibrium. Such reactions involve the specific binding of TFs to DNA sequences as well as the recruitment of RNA polymerase I complex. The dynamics of gene expression can be described by:

$$\frac{\mathrm{dmRNA(i,t)}}{\mathrm{dt}} = \mathrm{k_s} \prod_{\mathrm{j}} \mathrm{TFA(j,t)}^{\pi_{\mathrm{ij}}} - \mathrm{k_d} \prod_{\mathrm{j}} \mathrm{TFA(j,t)}^{\pi_{\mathrm{ij}}} \mathrm{mRNA(i,t)} \tag{5.1}$$

This power-law rate expression assumes a rate of synthesis depending on the activities of TFs whereas the degradation term is also considered proportional to the actual mRNA levels [97]. Making the quasi-steady state approximation for mRNA(t) and solving the corresponding algebraic equation leads to the following expression, accounting for an appropriate normalization with respect to the initial conditions:

$$\mathrm{mRNA(i,t)} = \frac{\mathrm{k_s}}{\mathrm{k_d}} \prod_{\mathrm{j}} \mathrm{TFA(j,t)}^{\pi_{\mathrm{ij}}} \Rightarrow \frac{\mathrm{mRNA(i,t)}}{\mathrm{mRNA(i,0)}} = \prod_{\mathrm{j}} \left[\frac{\mathrm{TFA(j,t)}}{\mathrm{TFA(j,0)}} \right]^{\pi_{\mathrm{ij}}} \tag{5.2}$$

A log-transformation results in the following generalized linear expression that relates the log-normalized

$$\mathrm{E} = \Pi \cdot \mathrm{P},\ \mathrm{E} = \log\left[\frac{\mathrm{mRNA(i,t)}}{\mathrm{mRNA(i,0)}} \right], \mathrm{P} = \log\left[\frac{\mathrm{TFA(j,t)}}{\mathrm{TFA(j,0)}} \right], \Pi = \left\{ \pi_{\mathrm{ij}} \right\} \tag{5.3}$$

where E matrix is the log-ratio of the gene expression level of gene i at time point t relative to the initial condition at t=0 ($\log \frac{E(i,t)}{E(i,0)}$) and its dimensions are N_g (number of genes) x N_T (number of time points), Π is the connectivity matrix whose entries are constant and characterize the strength of interaction (binding affinities) between any regulatory pair (i,j) with j to refer to the regulator and its dimensionality is N_g x N_{TF} (number of transcription factors) and P matrix contains the inferred effective dynamic activities for each regulator, expressed also as log-ratios, during time course of the experiment. Thus, its dimensionality is N_{TF} x N_T. The interaction strengths are either determined as an output of the decomposition [96] or assumed to be known and are proportional to the experimentally determined binding affinities of the transcription factor to the promoter region [93].

In our formulation we opted to treat the strength coefficients as surrogates for the binding affinity of the transcription factor to the promoter region in the sense that they should not be treated as condition dependent parameters but rather as a fundamental property of the system, since our primary motivation is to use as much of the available biological information as possible and minimize the amount of fitted parameters. The reconstructed activities will absorb any condition specific alterations to the transcriptional response. Therefore, the interaction coefficients will be considered to be either known from experimental studies [104, 105] or determined computationally by associating binding affinities to position weight matrices [106].

In addition to the strength of the interactions the directionality of the activation is also critical given that transcription factors are known to exhibit multifunction characteristics [107]. As a result, TFs are known to act as activators, repressors or exhibit both characteristics depending on conditions. Therefore, given the effective activity of a transcription factor we need to be able to simulate its corresponding effect, whether it is activating or repressing the expression of the

target genes. Assuming for simplicity that one TF regulates a single gene, then depending on the nature of the interaction the effect of changes in the TFA will have distinct effects on the changes in gene expression. If the activity of the factor increases and if the factor activates the expression of the gene, then the corresponding expression should increase. However, if the factor represses the expression of the gene, then the increase in activity should result in decrease in the expression of the gene. Equivalent arguments can be made for the case where the activity of the factor decreases.

We propose to model this by introducing a new variable, $P^{eff}(i,j,t)$ which represents the effective TFA for a given gene given that the type of interaction, either repressor or activator, has been identified. The definition is done through the introduction of a binary variable

$r(i,j) = \begin{cases} 1 & \text{TF(j) activates gene(i)} \\ 0 & \text{otherwise} \end{cases}$. Using r(i,j) and given the intrinsic activity of factor "j" we can now defined the effective activity which is a function of the pair (j,j) for each time "t":

$$P^{\text{eff}}(i,j,t) = \left[2r(i,j)-1\right] \cdot P(j,t) \tag{5.4}$$

The existence of an interaction element r(i,j) depends on the existence of a known regulatory interaction between factor "j" and gene "i" and the strength of the corresponding interaction will be assumed to proportional/equivalent to the binding strength.

Given, therefore, the architecture describing the superstructure of all possible regulatory interactions defined through the interaction matrix

$D(i,j) = \begin{cases} 1 & \text{TF(j) regulates gene(i), i.e. } \pi(i,j) \neq 0 \\ 0 & \text{otherwise, i.e. } \pi(i,j) = 0 \end{cases}$ we approximate the log-ratio of the expression data as:

$$E(i,t)=\sum_{j}\Pi(i,j)\cdot P^{eff}(i,j,t)+\text{error}$$
$$P^{eff}(i,j,t)=\left[2r(i,j)-1\right]\cdot P(j,t) \qquad (5.5)$$

The "error" term is incorporated to simulate error-in-measurement as well as other potential sources of uncertainty.

Analysis of regulatory networks

Deciphering the structure of regulatory networks should be considered as the prelude to further analyses that aim at elucidating putative roles of the regulators rather than a rigorous and restrictive reconstruction of experimental data. After all, it is widely accepted that multiple, alternative, regulatory networks can reproduce experimental data [97, 108]. As such, a number of questions emerge once a particular reconstruction has been determined, namely:

1. Can these networks be identified in a systematic and unbiased manner?
2. Are there any persistent interactions that emerge from multiple architectures?
3. Are there specific transcription factors whose activity profiles remain robust across multiple realizations?
4. Can the specific function of the undetermined factors, i.e., factors can act either as activators or repressors, be systematically determined?
5. Do preferential patterns emerge in terms of the nature, i.e., activator or repressor, of these factors be identified?

We are proposing a mixed-integer formulation able to effectively address all the aforementioned questions in a unified framework. The complexity of the regulatory network is controlled through the introduction of a binary variable z(j) which denotes the existence, z(j)=1, or non-

existence of a particular regulator z(j) = 0. The underlying assumption behind this modeling exercise is to identify what types of alternative structures can be constructed that reproduce optimally the experimental expression data. The complexity of the network is controlled by setting the required number of non-zero elements in this variable. Furthermore, alternative structures for the same number of transcription factors can be generating by introducing appropriate cuts that exclude previous integer solutions, i.e., combinations of non-zero z(j)'s [109].

It should be noted that the definition of P^{eff} introduces a non-convex bilinearity in the formulation due to the produce of the continuous variable P(j,t) and the binary variable z(j). However, this produce is exactly linearized through the introduction of the following set of constraints

$$\begin{aligned} &-r(i,j)\,M-P(j,t) \le P^{eff}(i,j,t) \le r(i,j)\,M-P(j,t) \\ &(r(i,j)-1)\,M+P(j,t) \le P^{eff}(i,j,t) \le (1-r(i,j))\,M+P(j,t) \end{aligned} \tag{5.6}$$

This set functions as follows: when r(i,j)=0 ("j" is a repressor of "i') the system reduces to:

$$\begin{aligned} &-P(j,t) \le P^{eff}(i,j,t) \le -P(j,t) \\ &-M+P(j,t) \le P^{eff}(i,j,t) \le M+P(j,t) \end{aligned} \tag{5.7}$$

Therefore, the second constraint is inactive (M is a big number) whereas the first constraint forces $P^{eff}(i,j,t)=-P(j,t)$. The implication is that because "j" acts a repressor of "i" if the activity of P(j,t) increases, i.e., P(j,t)>0, the effect of E(i,j,t) should be of the opposite sign and therefore result in reduction of E(i,j,t), i.e, E(i,j,t)<0. Similarly, if the activity of P(j,t) < 0, because "j" is a

repressor, then reduction in its activity should enhance the expression of E(i,j,t), i.e., E(i,j,t)>0.

When r(i,j)=1("j" is an activator of "i') the system reduces to:

$$\begin{aligned} &-M-P(j,t) \le P^{eff}(i,j,t) \le M-P(j,t) \\ &P(j,t) \le P^{eff}(i,j,t) \le P(j,t) \end{aligned} \tag{5.8}$$

This makes the first constraint redundant, whereas the second constraint forces $P^{eff}(i,j,t)=P(j,t)$ and therefore it acts as an activator.

The complete optimization formulation (miSARN) is as follows:

$$\min \sum_i \sum_t e^+(i,t) + e^-(i,t)$$

subject to

$$E(i,t) - \sum_j \pi(i,j) P^{eff}(i,j,t) = e^+(i,t) - e^-(i,t) \quad i = 1,...,N_g, \; t = 1,...,N_T$$

$$\sum_j z(j) = m \le N_{TF} \quad j = 1,...,N_{TF}$$

$$\sum_j D(i,j) \cdot z(j) \ge 1 \quad i = 1,...,N_g, \; j = 1,...,N_{TF}$$

$$-r(i,j) M - P(j,t) \le P^{eff}(i,j,t) \le r(i,j) M - P(j,t) \quad i = 1,...,N_g, \; t = 1,...,N_T, \; j = 1,...,N_{TF}$$

$$(r(i,j)-1) M + P(j,t) \le P^{eff}(i,j,t) \le (1-r(i,j)) M + P(j,t) \quad i = 1,...,N_g, \; t = 1,...,N_T, \; j = 1,...,N_T$$

$$z(j) P_{max} \le P(j,t) \le z(j) P_{min} \quad t = 1,...,N_T, \; j = 1,...,N_{TF}$$

$$\sum_{j \in N^k} z(j) - \sum_{j \in B^k} z(j) \le \left| N^k \right| - 1$$

$$N^k = \{j \mid z^k(j) = 1\}, \; B^k = \{j, \mid z^k(j) = 0\}$$

$$D(i,j) = \begin{cases} 1 & \pi(i,j) \neq 0 \\ 0 & \pi(i,j) = 0 \end{cases}$$

$$P(j,t), P^{eff}(i,j,t) \in \Re \quad i = 1,...,N_g, \; t = 1,...,N_T, \; j = 1,...,N_{TF}$$

$$e^+(i,t), e^-(i,t) \in \Re^+ \quad i = 1,...,N_g, \; t = 1,...,N_T$$

$$z(j), r(i,j) \in \{0,1\} \quad i = 1,...,N_g, \; j = 1,...,N_{TF}$$

N_{TF} *is* the number of transcription factors, N_g is the number of genes, and N_T is the number of time points.

In order to identify structurally robust elements of the regulatory architecture we introduce a robustness metric which quantifies the number of times a particular TF appears in each of the alternative structures in conjunction with the robustness of the reconstructed activity profile. The metric is therefore:

$$R(j) = \frac{f(j)}{Maxf} \times C(j) \tag{5.9}$$

Where R(j) is the robustness of TF "j" when we generate multiple network modules, f(j) describes the frequency of TF j across the multiple solutions (simply it shows how many times TF j appears in different network architectures), C(j) corresponds to the average Pearson's Correlation coefficient for the multiple inferred activities (P(j,t)) of TF j and finally is the total number of alternative structures under consideration.

Results

Experimental data

Temporal expression profiles of *E coli* during transition from glucose to acetate as the sole carbon source were detected by using DNA microarrays. The importance of such experiment lies on the premise that glucose and acetate are utilized by distinct metabolic pathways and thereby understanding such profiles in different carbon sources gives us a more thorough insight about the dynamic behavior of *E coli* [110]. The temporal *E coli* expression data as well as the connectivity matrix for this system are publicly available at http://ww.seas.ucla.edu/~liaoj/. The data included the log transformed expression levels (relative to initial time point) of 100 genes [99] recorded at 10 time points. Such expression data have been part of studies [93, 99, 111] . Taking these genes into account we identified the corresponding connectivity matrix given the available information of RegulonDB [102] database. The available connectivity information concerns 828 genes and 120 TFs coupled with the information whether a TF is known to inhibit or activate a specific gene or whether its regulatory role is unknown (activator or repressor. Removing genes that are not regulated by any TF as well as TFs that do not regulate a gene our dataset consists of 88 genes and 30 TFs. In a nutshell, our experimental data refer to 88 genes, 30 transcription factors (TFs) and 10 time points whilst the entries of $\pi(i, j)$ are s. Then, based on RegulonDB information we fix the binary variables r(i,j) to be either 0 or 1 if j is known to repress or activate gene I, respectively.

Systematic generation of alternative regulatory structures

The complete regulatory structure is composed of 30 transcription factors. Given the hard constraint that each gene must be regulated by at least on TF, the formulation miSARN becomes infeasible at m=18 since this many factors are need to guarantee that all genes are properly

regulated. Varying the control parameter m in the range of 18-30 TF generates an equivalent non inferior set as shown in **Figure 22**. Interestingly we observe that there are 5 different network architectures (for m=26 TFs up to m=30) that generate architectures resulting in the same reconstruction error, despite the fact that each utilizes a different number of TFs.

Given the availability of these alternative structures, we proceed to evaluate the robustness of each factor across the multiple solutions. The results are summarized in **Table 2**. It is clear that a critical subset emerges that not only persist as a selection of active TF, but also the corresponding reconstructed profiles are very robust across multiple solutions. The reconstructed profiles for all factors across all the 13 solutions (m=18,…, 30) are depicted in **Figure 23**. Associated with these multiple solutions is the fact that these solutions are able to reconstruct the observed mRNA gene expression level with a very high level of fidelity, and thus we have been able to establish at least the fact that our model is able to replicate the experimental results **Figure 24.**

TF name	relative connectivity	f(j)	C(j)	R(j)
Ada	1	13	1.0	1.0
CysB	4	13	1.0	1.0
FadR	3	13	1.0	1.0
GatR	4	13	1.0	1.0
LeuO	3	13	1.0	1.0
Lrp	6	13	1.0	1.0
PurR	3	13	1.0	1.0
TrpR	3	13	1.0	1.0
TyrR	6	13	1.0	1.0
ArcA	18	13	0.9	0.9
PhoB	5	13	0.9	0.9
FIS	7	11	1.0	0.9
NarL	9	13	0.9	0.9
CRP	21	13	0.9	0.9
RpoE	8	13	0.9	0.9
RpoS	5	13	0.7	0.7
FruR	7	13	0.7	0.7
OmpR	3	13	0.6	0.6
IHF	12	13	0.6	0.6
IclR	4	12	0.9	0.6
GlpR	1	8	1.0	0.5
LexA	1	5	1.0	0.4
PspF	1	5	0.6	0.4
FNR	16	10	0.4	0.2
CsgD	3	2	1.0	0.2
Rob	3	1	1.0	0.2
SdiA	1	5	0.2	0.1
RpoN	1	8	0.2	0.1
GalR	3	7	0.0	0.0
RcsAB	1	4	0.0	0.0

Table 2: The conservation associated with the different transcription factors used in our example.

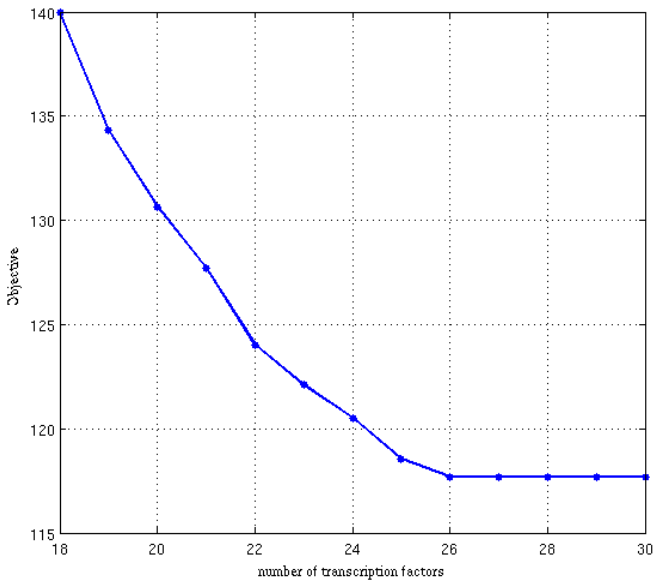


Figure 22: Objective value vs. size of regulatory network

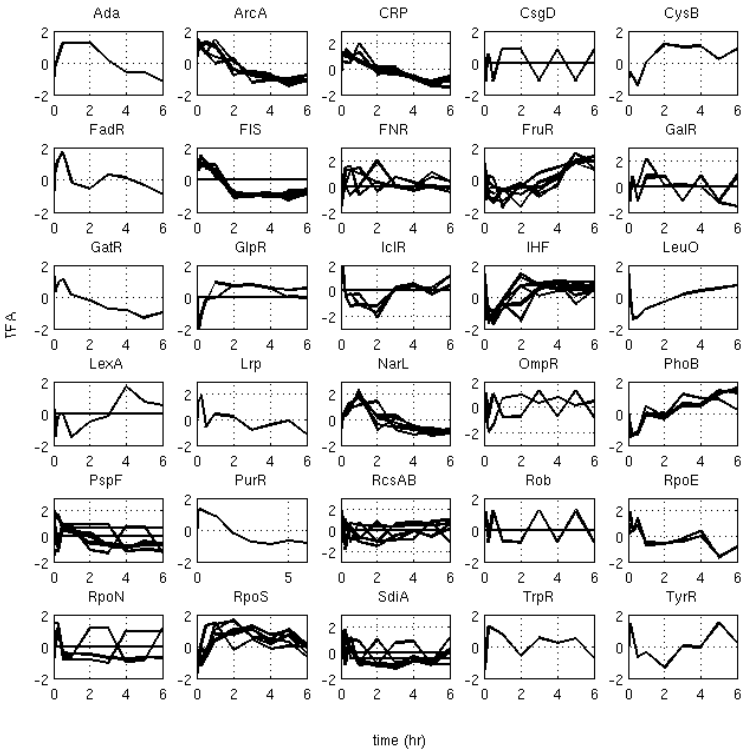


Figure 23: The dynamics associated with the transcription factors over multiple solutions

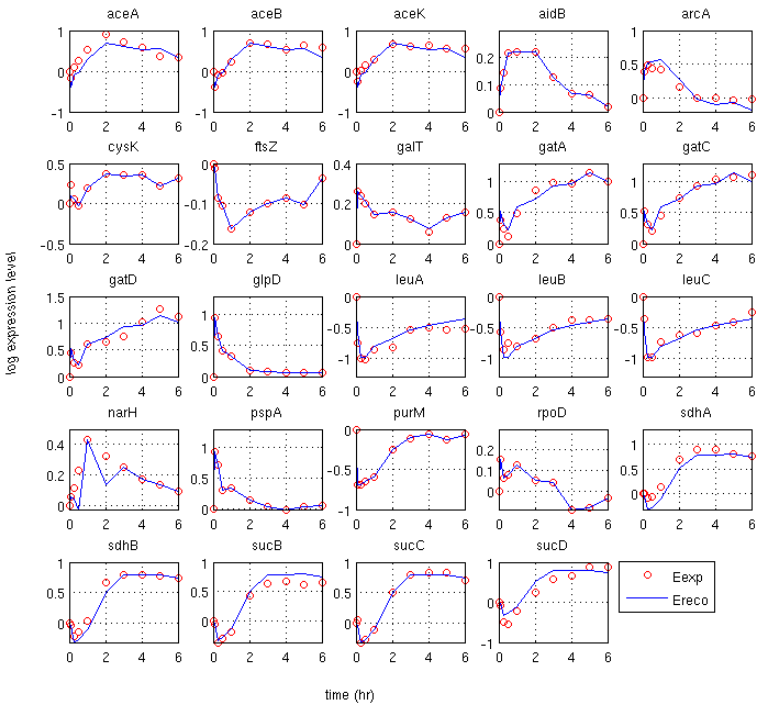


Figure 24: The dynamics associated with the reconstruction. As a sanity check, it is important to determine that our method is able to reconstruct the observed changes in gene expression

Discussion

One of the strengths of utilizing a MILP formulation is the fact that we are able to incorporate cuts into the solution. Thus, in addition to generating networks of different complexities, we are also able to generate related networks of the same complexity. This leads to the identification of multiple regulatory structures, of which interesting patterns emerge. The multiple architectures for m=29 effectively define networks in which one TF is eliminated from the network. The eight solutions are depicted in **Figure 25**. There are four distinct modules that give rise to these solutions and in all cases in effectively amounts to the elimination of factor provided that its contribution can be represented by another factor. The pairs that are exchanged in these eight solutions are: (PspF, RpoN), (SdiA, RcsAB), (CsgD, OmpR), and (Rob, GalR). These findings even though are merely computationally can be characterized as both challenging and promising on the premise that there is on-going research about identifying clinically intervention points whose effective combinatorial inhibition would improve the process of therapeutic drugs. There are several studies [91, 112] that are interested in unraveling the underlying principles that govern gene regulation by either combining sequence data with binding data such as Chip-chip data and expression data or by knocking out (deleting) transcription factors and binding sites with the aim to reveal more about functional regulatory interactions and pathways.

Based on the aforementioned results we can provide with a list of alternative network structures that will help us to understand the complex process of transcription regulation. Furthermore, when we solve for m<29 e.g. m=28 TFs, this means that we are deleting at each solution two TFs out of 30. The multiple solutions, which are 24, are subset of the four modules presented in **Figure 25** with the “rule” that two co-regulators that do not regulate any other gene at a neighbor module cannot be removed. All the other combinations are feasible (optimal) such as removing *CsgD* along with *Rob* or with *GalR* or with *RcsAB* and so on. Such “predictable”

solutions arouse to occur for *m* =27 and *m* =26. Thus, the four modules in **Figure 25** are the basis that accounts for all the solutions taken by the combinatorial removal of transcription factors given a predefined number of existing nodes (TFs) in the network. Overall, all these alternative structures may serve as a very useful tool in identifying combinations of regulators that can be knocked –out of the network. It is of major issue the biological validation of such results but in this study we are proposing the efficiency of our formulation to generate such multiple solutions with meaningful biological impact in modeling transcriptional regulatory networks and it is our future work to integrate such an algorithm with real biological examples.

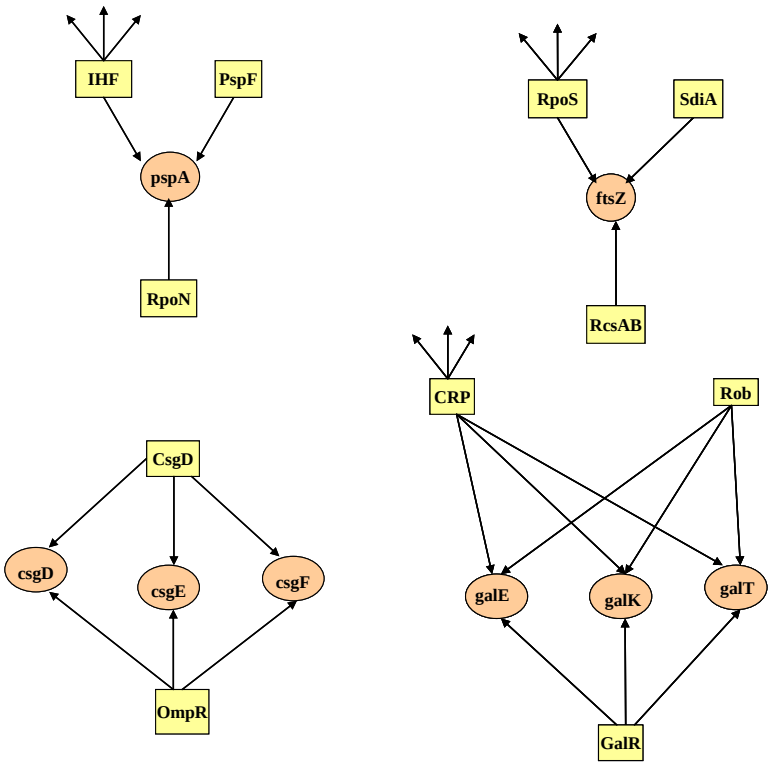


Figure 25: In these network motifs, it was found that the factors of similar connectivity can be eliminated without loss in the error function. For instance in the first figure (top right) PspF and RpoN represent interchangeable factors. The presence of such modules may signal the importance of a specific metabolic process in an organism due to the high level of redundancy in its regulation.

In addition to predicting the activity of a given transcription factor, the other goal of this work was to lower the ambiguity associated with transcription factor prediction. It was hoped that through the reconstruction of mRNA gene expression profiles, we can isolate the most probable connectivity structure required. Even though E coli, with well characterized transcriptional interactions were used, there was still significant ambiguity within the multiple regulatory structures. On one hand, the presence of this ambiguity is interesting because it provides a mechanism for silent mutations [113] i.e. deletions in the genome that appear to have very little effect upon the system. The results of our analysis suggest that there are some critical transcription factors which are necessary amongst all solutions, and these are exemplified by having very consistent profiles as shown in **Figure 23**. The transcription factors which are part of multiple alternative regulatory structures on the other hand have highly variable dynamics depending upon the network architecture.

These alternative regulatory structures however are determined via their binding interactions π, as well as the initial network which was fed into the problem. Thus, because the result of our transcription factor prediction, our network is not known with confidence, this bias is carried over to this formulation. Because of the impact of this original regulatory structure upon our solution, we do not hypothesize that this method can be used to further refine the results of our transcription factor prediction, unlike other methods such as Module Networks[114].

Most importantly is the fact that we are essentially solving a linear problem. Thus, the dynamics of our mRNA gene expression profiles represent linear combinations of transcription factor activity. Thus, whatever dynamics which were undertaken by our set of candidate genes will be a linear combination of the activity of transcription factors. This is problematic because of our previous result in which we isolated a small set of different dynamic motifs, whereas from the

results of the transcription factor prediction, we have obtained a much larger set of transcription factors. The lack of different patterns within our data as well as the ambiguity within our data prevents miSARN from being applied to our dataset.

Finally, given the linear nature of this problem, we have essentially shifted our question of how a dosing of corticosteroid affects a given mRNA gene expression level to the question of how a dosing of corticosteroid affects transcription factor activity. At this point, we have no way of linking our specific input into the problem. Thus, while the method may provide useful insight into the most probable regulatory path taken in a given phenomenon, it does not answer the primary questions which we had when we began the analysis.

Conclusion

One of the interesting phenomena which we were able to observe in our evaluation with *E coli*, was the presence of alternative networks in which the reconstruction error remained essentially constant. What was more interesting in these results was the fact that these alternative regulatory structures appear to correspond to experiments in which gene knockout strains do not show significant change in the viability of the organism. Despite the initially promising results which we have obtained utilizing *E coli.,* we were unable to further extend the method to our system due to ambiguities within the original network which is an input into this formulation. Secondly, because we had no method for incorporating specific inputs into the system, we were unable to link our specific inputs to transcription factor activity. The primary confounding factor in this formulation is the fact that we are trying to determine too many unknowns in the problem, specifically the transcription factor activity as well as the connectivity structure. This is problematic because of the interdependence one factor has upon the other. As

seen in the cases when we look at multiple solutions, there exist many cases where alternative regulatory structure yield significant differences within transcription factor activity.

However, while this analysis may not be useful for identifying specifically the regulatory elements within our system, it was useful in framing our problem better. In order to determine the underlying regulatory interactions that form the basis of our mechanism, it is necessary to either know the specific active transcription factor network, or the activity or relevant transcription factors. Secondly along with the need to obtain transcription factor activities, it is also necessary to be able to incorporate the dynamic input of our system into the model. By doing so, we ought to be able to obtain better insights as to how a stimulus in the form of drug administration is able to affect either mRNA gene expression level or the activity of the transcription factors. Therefore, rather than basing the rest of our analysis upon only mRNA gene expression data and trying to solve for the connectivity structure as well as transcription factor activity, we will instead by focusing upon an experimental system that allows us to directly measure a surrogate for transcription factor activity.

Equation Chapter (Next) Section 1Obtaining Dynamics of the Glucocorticosteroid Receptor via the Living Cell Array

From the prior analysis of the temporal gene expression data, we have obtained the various pieces of information. The primary driving force behind the response to a bolus injection of corticosteroids is the endogenous concentration of the drug. However, under the chronic administration of corticosteroids, we see a significantly different result. In the chronic case, we see that the genes related to metabolism appear to follow the same drug concentration mediated response, but the genes related to inflammation and the immune response do not. Rather than being directly related to the concentration of the drug, the response decreases back to baseline despite continued drug administration. Because of this difference in the dynamic response, it is difficult to rationalize that this response comes from the same underlying mechanism. While intuitively, the understanding is that because we are merely administering a drug in a different manner, there should be one mechanism which is active over the different dosing strategies that give rise to these disparate responses.

However, this mechanism could take several different forms. For instance, one possible hypothesis is that the chronic administration of corticosteroids may lead to a build-up of intermediate metabolites, which triggers a secondary mechanism that differentiates the response of the organism under the chronic administration of corticosteroid as compared to the acute administration of corticosteroids. Other hypotheses as to the underlying response may be the existence of an orphan receptor which is sensitive to corticosteroid administration, or that corticosteroids may have significant non-transcriptional activity. At this point, there are a large number of possible hypotheses that can be used to explain the observed transcriptional dynamics aside from the ones that have been enumerated here. Therefore, testing all of the

possible hypotheses experimentally in a reasonable amount of time is not feasible. The question is however, whether we can use the concepts behind systems biology to eliminate most of these hypotheses and generate a smaller set which can then be tackled experimentally.

Central to these hypotheses is the activity of the corticosteroid receptor, and its relation to the activity of genes related to metabolism or inflammation. The common thread that ties these disparate hypotheses together is the fact that how the activated corticosteroid receptor responds is unknown. Therefore, if it were possible to first identify the dynamics of the corticosteroid receptor as well as the transcription factors related to inflammation or metabolism, it would go a long way to allowing us to eliminate the set of initial hypotheses.

Aside from the failure to determine the presence of common regulators in our extracted genes, it is also important to note that the activity of a transcription factor is not determined entirely by the amount of transcription factor in the system. Therefore, we cannot use the gene expression of the corticosteroid receptor as a surrogate for activity[115]. This is because while the amount of the receptor in the system plays a key role in the activity level, we must account for factors such as activation via a ligand, dimerization, or phosphorylation[116]. Therefore, even though a given transcription factor may be over-expressed, its activity may be negligible if the activating factor is not present.

There exist numerous methods for assessing transcription factor activity. Experimental techniques such as ELISA[117] attempt to quantify the binding between the activated transcription factor with their consensus sequence, and computational methods such as NCA[96], PLS[93], NIR[118], and Module Networks[114] attempt to de-convolve transcription factor activity based upon either known or hypothesized binding interactions as well as mRNA

gene expression data. These computational techniques are generally categorized as network reconstruction techniques.

However, in the context of our work, all of these techniques have significant limitations. For instance ELISA, as an offline experimental technique does not offer the time resolution we need for model building. Furthermore, because this technique measures the activity level of only a single transcription factor, we are unable to apply the concepts of systems biology, specifically the analysis of how the different systems interact i.e. how the activation of corticosteroid plays a role in the regulation of other transcription factors.

The computational methods such as NCA, PLS, and Module Networks are limited because they rely upon transcription factor binding data, which we do not have a comprehensive and validated set, the same factor which confounded our analysis utilizing miSARN. Therefore, while it should be possible for us to utilize transcription factor prediction algorithms to compensate for the lack of experimental data, the ambiguities associated with the binding site predictions creates another level of ambiguity as to the results. Additionally, these algorithms are required to make a large set of assumptions as to the underlying structure of their networks because the problems are ill-posed[1]. These constraints upon the network architecture may not necessarily be reflected in the underlying biology and therefore may add an additional level of uncertainty to our conclusions. Finally, as we have seen before in our evaluation of the miSARN methodology, these methods do not translate the input in the form of the drug stimuli into the output which is the mRNA expression levels. What these methods have done is moved the question of how does the drug impact mRNA levels, to the question how does the drug impact transcription factor

[1] Ill posed is defined as having more unknown variables than equations, and is not a criticism of the underlying method, formulation, or hypothesis

activity. Therefore, what is truly needed is a method that allows us to measure the dynamic activity of a transcription under stimulation.

Ideally, we need a method for first measuring the levels of multiple transcription factors, preferably related to inflammation, the immune response, or metabolism under high temporal resolution. Furthermore, these need to be measured in such a way that the model we define will not be ill-posed[1] without making any assumptions as to the architecture of the network. Fortunately, such a system does exist in the form of the Living Cell Array. The experimental design of the Living Cell Array allows us to assess the dynamics of the corticosteroid receptor as well as its effect upon the transcription factors related to inflammation. While it would have been beneficial to also assess the response of other factors related to metabolism, our current hypothesis is that the current indirect effect model is sufficient to replicate the dynamics under both dosing strategies.

The Living Cell Array (LCA) [119, 120] presents a unique experimental platform that allows for the direct estimation of the activity of a transcription factor. Rather than focusing upon the binding of transcription factors, or mRNA expression changes, it utilizes fluorescent reporters that respond to the levels of active transcription factor activities through monitoring the expression of a reporter GFP. Given the novel nature of its experimental design, the LCA offers the opportunity to decipher mechanisms driving the cross-activation of assemblies of transcription factors. Furthermore, owing to the design of the microfludics device, one can simultaneously obtain the levels of transcription factor activity under multiple stimuli with high temporal resolution. This greatly improves our ability to decipher the interactions between the different transcription factors because we are no longer constrained by limitations in the data,

where the number of genes measured is much greater than the number of conditions or time points in which they are measured [121].

We propose the creation of two methodologies for the analysis of the Living Cell Array. The first method of analyzing the Living Cell Array involves the use of a bi-clustering formulation[122]. This approach utilizes the co-expression of the various reporters to create a network which hypothesizes the direct interactions between two different transcription factors. The second method for identifying the network is referred to as Reverse Euler Decomposition (RED) which functions as a computational framework enabling us to quantify the dynamics associated with the connections. Combining this computational framework along with the experimental system of the Living Cell Array, it is possible to not only isolate TF interactions but also to quantify numerically evidence of nonlinear phenomena that are present in biological systems [123].

Superficially, the difference between the bi-clustering formulism and the RED formulism lies in the fact that the RED formulism allows for both the identification of regulatory dynamics in addition to the network architecture along, whereas the bi-clustering formulation is only able to identify the network architecture. However, these two alternative strategies represent a trade-off between accuracy and analysis time, with the RED formulism allowing for a more accurate determination of network architecture and dynamical reconstruction and the bi-clustering analysis taking much less time. However, in addition to the superficial difference in the algorithms, we hypothesize that because the networks obtained via the two different techniques should be closely related, the small differences in the networks may provide additional insight as to the underlying architecture of the system specifically those relating to “silent mutations.”

Transcription Factor	Stimulus
NFkB	TNF-α
AP1	IL-1
STAT3	IL6
ISRE	IFN-γ
GRE	Dexamethasone
HSE	No Direct stimulation

Table 3: The measured transcription factors and their associated stimulus

The Living Cell Array

The Living Cell Array is a microfluidics device which utilizes cells transfected with reporter plasmids. These reporter plasmids comprise of an unstable green fluorescent protein (GFP), a minimal promoter, and 4 repeats of a transcription factor's consensus sequence [124]. Therefore, when a transcription factor is in its active state, it binds to the plasmid thereby causing the synthesis of an unstable GFP [125]. In this system, the fluorescence levels act as a surrogate for the amount of activated transcription factor present within the system, because due to the artificial construction of these plasmids, fluorescence level of a given reporter should be determined only through the activated level of its associated transcription factor. However, it was found that under multiple stimulation profiles, there was a significant level of cross talk. We hypothesize that such cross-talk is due to interactions between the different transcription factors i.e. the activation of transcription factor A can cause the up/down regulation in the activity of transcription factor B.

Guided by the interest in hepatic inflammation the reporter cell lines were designed to probe the dynamics of transcription factors associated with inflammation [126]. Appropriate soluble stimuli were designed hat stimulate the dynamic cellular microenvironment and would enable the systematic characterization of the cellular responses. Specifically, the NFκβ transcription factor was induced by TNF-α, AP-1 induced by IL-1, STAT3 induced by IL-6, ISRE induced by INF-γ, GRE induced by Dexamethasone, and HSE which was not directly simulated **Table 3**.

Reconstructing Network Interactions from co-expressed Reporters

Bi-clustering or condition specific clustering attempts to isolate genes that are co-expressed under a specific set of conditions[127]. The hypothesis behind the utilization of this method is

that transcription factors which show very similar responses under multiple conditions are likely to be closely coupled, and thus related. Bi-clustering is nominally performed over a set of genes vs. conditions with only a single value per condition. However, in the given dataset, each gene/condition combination is described as a time series. In bi-clustering, genes that have similar expression values under a given condition are considered as possible candidates to be clustered together for that specific condition. Given the temporal expression data, the temporal expression can be simplified into an integer, so that gene expression profiles with the same integer would have similar expression profiles. This could have been accomplished in a variety of ways from Hashing Based methods[128], to standard clustering algorithms in which the cluster memberships are used to assign an integer denoting similarities in the expression profiles of different genes under a given condition. Thus as in the case of hash based clustering, genes denoted by a similar integer will have similar responses.

For this problem k-means clustering with a cosine similarity metric[129] was selected. K-means was run with 4 clusters, the minimum number of clusters needed for consistent clusters over multiple runs. Therefore, the temporal expression profiles were converted into integers that indicate the similarity under a given condition of 2 or more genes. K-means was chosen over a more sophisticated method such as HOT SAX because the small size of this problem did not necessitate the increase in speed, or the evaluation as to whether the system underwent significant perturbations.

Bi-clustering itself is NP-Hard[130], and therefore most of the algorithms which have been used for bi-clustering are heuristics[127]. The most obvious problem with most of the techniques which are based upon heuristics is the fact that they do not solve the problem to optimality. However, the more glaring problem is the inability for most of the heuristic based methods to

identify an arbitrary number of over-lapping bi-clusters. In most of the bi-clustering algorithms, finding multiple solutions involves removing a previously found bi-cluster from the dataset through techniques such as setting all of the values in a previous found bi-cluster to random numbers therefore breaking up any relationships within that cluster. There has been some work to find overlapping clusters[131]. However, such techniques are limited in the fact that one must determine before the structure of the overlap such as overlapping percentage as well as the number of possible overlapping structures within the data, something which is not known *a priori*.

The issue of overlapping bi-clusters is important because with non-overlapping bi-clusters, the networks which can be reconstructed from expression data will be a set of disjoint and independent networks. This contradicts with the general notion that transcriptional networks form highly interconnected networks [88]. Therefore, networks generated from the current algorithms cannot fully capture the level of interconnectedness. The advantage of utilizing a math programming approach is that it is very easy to exclude previous solutions and re-solve the problem to find other bi-clusters which may overlap with a previous solution. Without overlapping bi-clusters, the overall network is then reduced to a set of independent cliques of which the most complex network which can be created is a feed forward network**.**

The biggest issue that complicates the search of overlapping clusters is illustrated in

Figure 26. The primary problem is that after an optimal solution is found and that solution is rejected, there exists an overlapping cluster which is wholly a subset of the original solution. A mixed integer optimizations framework was selected due its ability to explicitly model constraints as well as solve the problem to global optimality, something which cannot be

guaranteed with the standard heuristic based method. In this mixed-integer framework, it is possible to eliminate a solution as well as all subsets of its solution through a modified system of integer cuts.

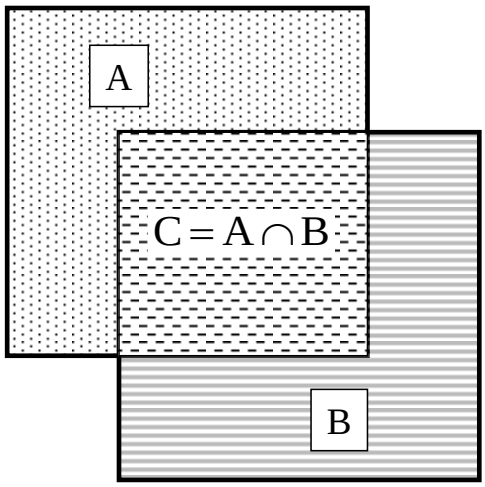


Figure 26: The problem of overlapping bi-clusters: Given two bi-clusters, A and B, the intersection of the two bi-clusters, C should be eliminated

One of the issues with using a formal mixed integer formulation is that it requires solving the full problem and not conducting an approximation. Therefore, the issue with the problem being NP-hard still remains. The mixed integer formulation mitigates the problem efficiently through intelligent pruning of infeasible and sub-optimal solutions, but does not change the overall algorithmic complexity. In the current iteration of the LCA, there are 6 specific transcription factors being utilized under 6 different conditions, and therefore the computational complexity is not an issue. However, in the most comprehensive case for transcriptional regulation, the problem set is still relatively small, on the order of 200 transcription factor binding sites having been quantified[104], and therefore still within the limits of solvability.

The mixed integer formulation is divided up into two portions, the bi-clustering formulation (6.1) , and the subset removal cuts (6.2). The problem is solved parametrically for the number of genes starting from N genes and decreasing until the number of genes equals 2. The optimization criterion maximizes the number of conditions. With this formulation, it is not necessary to define constraints of what a good bi-cluster entails though such constraints could be formulated. We find this to be an artificial constraint, for there could very well exist two genes which are well correlated over a large number of different conditions, of which the implications would be just as important as a bi-cluster of 10 genes that were well correlated over fewer conditions.

$$[(\lambda_i + \lambda_j + \mu_k) - 3] * M \le (\lambda_i + \mu_k) * D(i,k) - (\lambda_j + \mu_k) * D(j,k)$$
$$[3 - (\lambda_i + \lambda_j + \mu_k)] * M \ge (\lambda_i + \mu_k) * D(i,k) - (\lambda_j + \mu_k) * D(j,k) \quad (6.1)$$

The bi-clustering portion described in (6.1) requires the discretization of the signal. This works well for the time series data which is provided by the LCA. It essentially checks to see if two genes under a given condition have the same value with binary variables to indicate whether a given gene is included for the assessment. In (6.1), D represents the integer transformed data, λ represents the genes selected within the bi-clusters where μ represents the conditions under which the genes are co-expressed. The indices i,j,k represent the index in the array for which the gene or condition exists. M represents a large number that functions to essentially eliminate the constraint either of the two genes or conditions are not part of a given bi-cluster. In other words, genes i and j belong to bicluster k, i.e., $\lambda_i = \lambda_j = \mu_k = 1$, if and only if the symbolic representation of both genes are the same under condition k, i.e., D(i,k) = D(j,k). This is the only situation that would make (6.1) feasible. If $\lambda_i = \lambda_j = \mu_k = 1$ whereas $D(i,k) \neq D(j,k)$ (6.1) would be infeasible since the left hand side of both inequalities will be zero, whereas the right hand side is not. A schematic of how this assessment finds bi-clusters is shown in **Figure 27.** In ***Figure 27*** there are two λ variables which denote the two genes which are being checked for co-expression whilst the μ represents the condition in which they are checked from. If two genes are part of a bi-cluster, then the value under the two different conditions ought to be identical.

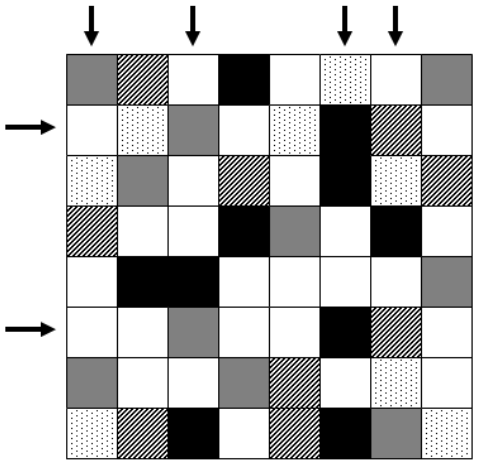


Figure 27: A schematic of how the formulation in Equation 1 works. Rows indicate genes, columns indicate conditions. Two genes (λ_2 = 1 and λ_6 = 1) are similarly expressed under four condition (μ_k = 1, k=1, 3, 6, and 7).

The problem with excluding subsets is simplified by the fact that the problem will be solved to optimality at every iteration and parametrically solving for different number of genes. The primary idea behind (6.2) is that a new solution requires a condition to be included that was not in a previous solution. (6.2) guarantees that each solution will not be a subset of a previously identified set of conditions. In (6.2), μ_k^{iter} represents the previous solution and μ_k^{citer} represents the current solution which may or may not be excluded. Therefore, the biclusters are generated sequentially and the exclusion constraints of (6.2) guarantee that the bicluster at iteration "citer" is not a subset or the previous clusters "iter".

Figure *28* illustrates how the subset removal cuts works. Equation (6.2) essentially forces the next possible solution to include a condition that was not included in a previous solution. If the current solution is a subset of any previous solution, then the following holds.

$$\sum_{P(iter)} \mu_k^{citer} < \sum_{k} \mu_k^{citer} \quad \forall iter < citer$$
$$P(iter) = \{i \mid \mu_k^{iter} = 1\} \tag{6.2}$$

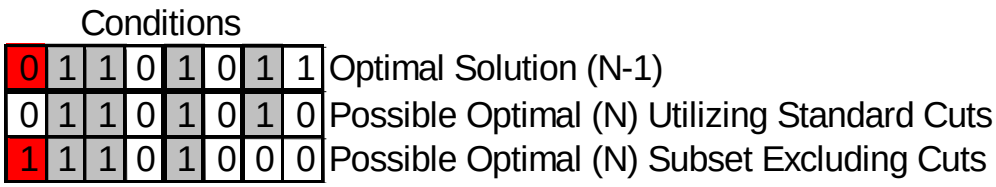


Figure 28: The solution for iterate (N-1) has 5 conditions, the next optimal solution has 4. However, the solution which is wholly a subset of a previous solution should be excluded.

Given that the formulation solves for the maximum number of condition under which N genes is co-expressed, the exclusion only occurs for the set of conditions. The set of cuts can be limited to only the conditions rather than the genes because the problem is solved parametrically with the maximum number of genes which should give the smallest number of conditions which these genes are co-expressed under. Once the number of genes has been decreased, the set of conditions in which the genes are co-expressed ought to have at least one condition which was not present in the previous solution. Therefore, by solving it parametrically, in N it is possible it removes the complexity of requiring a subset excluding cut from requiring both the conditions as well as the set of genes greatly simplifying the formulation.

After the bi-clusters were generated, they were evaluated as to whether or not one of the condition/reporter interactions in that bi-cluster had a 2-fold change in the overall activity. The data was reported in fold-change, and it was found that in the negative control case, the variability in the overall intensity differed by less than 2 fold. We opted to select bi-clusters which had at least one of the condition/reporters show a twofold change instead of filtering out the gene/condition combinations and then conduct the bi-clustering because it represented a compromise between focusing solely upon co-expression or the intensity values. The overall formulation is given in (6.3)**.**

$$\max \sum_{k} \mu_k^{citer}$$

$$s.t \sum_{i} \lambda_i^{citer} = N$$

$$[(\lambda_i^{citer} + \lambda_j^{citer} + \mu_k^{citer}) - 3] * M \le (\lambda_i^{citer} + \mu_k^{citer}) * D(i,k) - (\lambda_j^{citer} + \mu_k^{citer}) * D(j,k)$$

$$[3 - (\lambda_i^{citer} + \lambda_j^{citer} + \mu_k^{citer})] * M \ge (\lambda_i^{citer} + \mu_k^{citer}) * D(i,k) - (\lambda_j^{citer} + \mu_k^{citer}) * D(j,k)$$

$$\sum_{P(iter)} \mu_k^{citer} < \sum_{k} \mu_k^{citer} \quad \forall iter < citer \qquad (6.3)$$

$$P(iter) = \{i \mid \mu_k^{iter} = 1\}$$

D(i,k)= symbolic representaion of gene "i" in condition "k"

$$\lambda_i^{citer} = \begin{cases} 1, \text{ if gene i belongs to bicluster "citer"} \\ 0, \text{ otherwise} \end{cases}$$

P(iter), Q(iter) = denote the set of conditions that comprised previous biclusters

Network Reconstruction of the Bi-clustering Results

The result of running the bi-clustering formulation upon the data is a bi-partite network, in which there is an input layer which consists of the different stimuli into the system, and an output layer which consists of the measured transcription factor. However, it is possible to convert this bi-partite representation into that of a directed graph, thus making the result amenable for use in RED.

The primary purpose behind bi-clustering was to construct a network which gives insight as to the underlying mechanism which gave rise to the observed responses. Without any *a priori* information, a bi-partite network could be obtained in which links can be created from a regulator to a set of genes, if those regulators and genes are found in the same bi-cluster**.** However by incorporating additional information which is available due to the artificial construction of the reporter genes, one can generalize the bi-partite graph into a directed graph which gives insight as to the signaling cascade, specifically in this case, the induction of inflammatory/anti-inflammatory signals via external stimulus.

The specific piece of information which is utilized is the fact that the reporter genes can only be activated by their specific transcription factor, and therefore the only direct links that can be present in the graph is from a transcription factor and its specific reporter. These direct links are given in **Table 3** of the original LCA manuscript. [119]. Thus, if a bi-clustered determined that STAT3 and NFkB were co-expressed under a stimulation of TNF-α, then could utilize **Table 3** to hypothesize that the activation of STAT3 by TNF-α occurs via an NFkB intermediate.

Deconvolution of Network Interactions

The primary hypothesis behind the Deconvolution of Network Interactions that a given network architecture is defined not only by the links between the nodes, but also the dynamical response of these interactions. Thus, we propose the creation of a method which evaluates the dynamics of network architectures. By combining the network architecture, and these dynamical responses, it should be possible to reconstruct the experimental data in an accurate manner. Furthermore, because the reconstruction can be compared to the experimental data, it is also possible that such a method can also to assess the accuracy of a given architecture, and thereby search for an optimal network structure.

This is possible because the construction of the LCA experiment allows us to monitor the temporal dynamics of a system of TFs as they respond to a continuous infusion of soluble signals designed to activate specific TFs. In a hypothetical scenario only one factor should be activated for a given infusion of its corresponding activation signal. However, due to the cross-talk between TFs indirect interactions emerge which manifest themselves through the coordinated activation of an ensemble of factors. In order to decipher the emerging dynamic of the network of interacting TFs we need to first define an appropriate model for the dynamics of the system.

In its most general form, the dynamics of the system can be described as:

$$\frac{dTFA(i,t)}{dt} = F(TFA(j,t), j = 1,\ldots,N_{TF}) + s(i) \quad \forall t; i = 1,\ldots,N_{TF} \tag{6.4}$$

where N_{TF} denotes the total number of TFs in the network and TFA(i,t) represents the activity of transcription factor i, F represents an arbitrary function which incorporates and convolutes the underlying dynamics of the interacting TFs . The component s(i) expresses the effect of the activation event of a transcription factor. In the context of the LCA design it corresponds to a constant infusion of a soluble factor activating the TF and it is considered to be the known external stimulus that activates the transcriptional machinery. Essentially, in this model, we suggest that the dynamics of transcription factor activity can be described through an appropriate, yet to be determined, function (F) which is dependent upon the activity itself, and a forcing function, s, which may or may not be a function of time indicating a specific and direct activation of a transcription factor. In the context of LCA the forcing function is assumed to be independent of time since it is presumed that the soluble factors continuously activate the TFs through infusion. The activity of the TFs is quantified through the monitoring of the expression of the corresponding reporter genes.

A widely used simplification [132] approximates (6.4) as:

$$\frac{dTFA(i,t)}{dt} = \left[\sum_{j=1}^{N_{TF}} f(i,j,t)\cdot TFA(j,t)\right] + s(i) \quad \forall t; i, j = 1,\ldots,N_{TF} \tag{6.5}$$

This transformation effectively makes use of the assumption that the effect of the network of interacting TFs is additive and therefore the driving dynamics, as defined by the function F(TFA), can be decomposed in to $\sum_{j=1}^{N_{TF}} f(i,j,t)\cdot TFA(j,t)$ Underlying the transformation is the hypothesis

that the transcription factors do not form significant interacting complexes, and that transcription factors interact with each other independently [133]. Furthermore, the model assumes a connection weight that maps the influence of one transcription factor to another [93, 99]. Thus, the, yet to be determined, function f(i,j,t) describes the influence of TF i to the activity of TF j at time t.

Several methods have been proposed that solve for the functions f(i,j,t) . A commonly invoked assumption is that the interaction strength, quantified through f(i,j,t), is not a function of time [118, 134, 135]. This treats the interactions as scalars representing effectively the network connectivity strength. For instance the Network Identification by Multiple Regression, NIR, [118] assumes that the transcriptional dynamics are measured at steady state, and therefore eliminate the contribution of time upon f(i,j,t) whereas other methods, such as Dasika et al. [135] and Schmitt et al. [136], use time delays as a method of identifying when in time there exists a significant interaction, thus removing the explicit temporal modeling as well.

Given the lack of sufficient conditions in the experimental data, most algorithms must also account for the fact that using the available experimental data, the problem is ill defined i.e. there are more variables than equations in the formulation. As a result, numerous ingenuous approaches have been proposed that make use of innovative ideas to overcome such limitations. In that respect NIR constrains the number of allowed connections for to the number of conditions measured [118], whereas Guthke et al. use Singular Value Decomposition (SVD) to reduce the number of genes whose profiles need to be reconstructed. Network Component Analysis (NCA) [96] rigorously defines the number of active interactions which can be present. By gauging the effect of unmeasured transcription factors have upon gene expression profiles, NCA establishes a set of related connectivity structures such that the solutions differ by a diagonal

scaling matrix. However, it should be noted that NCA does account for the temporal evolution of the interaction strengths.

However, when TFA(i,t) is known at a relatively high temporal resolution one could in principle argue that a numerical estimate of the interaction dynamics, as expressed by f(i,j,t) can be obtained. In the case of a single variable – single equation, the system is fully determined at each time point. Therefore, if the dynamics of $\dot{x} = f(x)$ and represented via the decomposition dx/dt = α(t) x(t), it is possible to determine α(t) in a numerical sense provided that both dx/dt and x are known. This is done simply by assuming that the dynamics expressed as can be resolved at each time point by simply evaluating α(t) as:

$$\alpha(t) = \frac{\left(dx/dt\right)_{estimated}}{x(t)_{measured}} \tag{6.6}$$

This process is essentially the reverse of Euler integration in which a(t) and x(0) are initially known and we wish to reconstruct the dynamics of x(t) in a numerical sense. The aforementioned calculation assumes an accurate estimate of the rate of change of x(t) based on the measured values of x(t). Because we assume that at each time point a single parameter needs to be determined, α(t), then the system is fully determined and assuming that the operation in (6.6) is possible the instantaneous dynamics can be resolved for. However, with N_{TF} transcription factors measured under a single stimulus, at each time point there are effectively N_{TF}^2 unknowns, f(i,j,t), $\forall$t, since each transcription factor may be interacting with every other transcription factor. To fully account for these unknowns, it is necessary to evaluate the set of differential equations under at least N_{TF} different starting points or different conditions for the problem to be fully defined. The LCA framework allows for the concurrent definition of such multiple experimental perturbations by the introduction of either independent soluble signals,

or combinations of such signals in an effort to activate groups of TFs simultaneously. What makes this possible is the unique data generated from the LCA. The key advantage of utilizing the LCA is that for each transcription factor measured, it is reasonably straightforward to add one or more conditions such that the system is fully defined. This is because each condition represents the stimulation of the system with a stimulatory soluble factor, denoted earlier by s(i). It is important to note that in both our formulation as well as the experimental system, multiple combinations of soluble factors can be utilized as separate conditions. Therefore, no simplifying assumptions need to be made regarding the complexity of the network.

We refer to the process of generating an approximation to f(i,j,t) as Reverse Euler Decomposition, (RED). In a similar fashion to Euler Integration, we seek a numerical solution to the problem. However instead of defining the problem as finding a numerical representation of the response, TFA(i,t),as in the case of Euler Integration, we shall be looking for a numerical representation for f(i,j,t), which is normally known analytically in Euler integration but unknown in our case, and hence the moniker Reverse Euler Decomposition.

Thus the purpose of RED is to evaluate numerically the interactions dynamics, f(i,j,t) at each time point. Given the available experimental data, we are effectively performing a least squares estimation at each time point through the minimization of an appropriate norm:

$$\left\| \left.\frac{\mathrm{dTFA(i,t)}}{\mathrm{dt}}\right|_{\text{estimated}} - \left(\left[\sum_{j=1}^{N_{TF}} \mathrm{f(i,j,t)} \cdot \mathrm{TFA(j,t)} \right|_{\text{measured}} \right] + \mathrm{s(i)} \right) \right\| \qquad \forall t; i, j = 1, \ldots, N_{TF} \tag{6.7}$$

Furthermore, given the time resolution, the derivative of each transcription factor's activity level can be accurately estimated via smoothing splines [137]. Thus, the rate of change of TFA(i,t) is numerically estimated given the measurements of TFA(i,t). The minimization of the norm essentially minimizes the error between the rate of change of TFA as measured from the data

and the rate of change in TFA as predicted by the model. In this formulation f(i,j,t), represents the contribution of one transcription factor upon the activity of another TF at any given time point.

However, from an analysis point of view a critical question which emerges is whether the network of interacting TFs possesses any special structural characteristics. In other words, whether the network is composed of fully interacting elements, or whether direct links between specific TFs do not exist. These would effectively be translated to:

$$\exists(i,j) \ni : f(i,j,t) = 0 \ \forall t \tag{6.8}$$

In order to address this question, we will couple the deconvolution of the dynamics, based on the minimization of [4], with mathematical programming formulations that allow for the optimal identification of the network architecture, i.e., direct links between TFs, as well as the deconvolution of the network dynamics. In fact we present two modeling approaches, one which optimally determines interactions, and a second formulation which utilizes an *a priori* network architecture. This *a priori* network architecture may be the result of other analysis such as prediction algorithms for transcription factor binding [138], chip-chip experiments[105], or other algorithms such as Boolean networks which link the activity of a given gene with its particular activator [139].

Global Network Reconstruction via Reverse Euler Deconvolution

The LCA provides the opportunity to generate multiple realizations of the TFA dynamics based on the multiple systemic perturbations through the infusion of soluble factors activating the target TFs. In order to explore the wealth of the data and to extract what would appear to be the underlying interaction dynamics representative of the systemic response across a number of

conditions, we deconvolute simultaneously, at each time point, all the experimentally generated profiles. Therefore, at each time point a number of conditions, equal to the number of TFs in the system, are used for the estimation of the dynamics. In order to render the problem computationally tractable and maintain a linear nature, we opt to utilize the L-1 norm as opposed to the more widely used L-2 norm:

$$\left| \frac{dTFA(i,t)}{dt}\Big|_{estimated} - \left(\left[\sum_{j=1}^{N_{TF}} f(i,j,t) \cdot TFA(j,t)\Big|_{measured} \right] + s(i) \right) \right| = \epsilon(i,t) \quad \forall t; i = 1,\ldots,N_{TF} \tag{6.9}$$

The global network reconstruction formulation simultaneously attempts to identify the most probable network architecture i.e. a network architecture which yields the lowest error as well as the numerical solution for f(i,j,t). Therefore, the network reconstruction optimization problem reconciling the dynamics over a number of external disturbances, k, is defined as follows:

$$\min : \sum_{i=1}^{N_{TF}} \sum_{k=1}^{N_c} \sum_{t=1}^{T} \epsilon^{+}(i,k,t) + \epsilon^{-}(i,k,t)$$

subject to:

$$\left.\frac{dTFA(i,k,t)}{dt}\right|_{estimated} - \left(\sum_{j=1}^{N_{TF}} \left[f(i,j,t) \cdot TFA(j,k,t)\big|_{measured} + \beta(i,j) \cdot s(j,k) \right] \right) = \epsilon^{+}(i,k,t) - \epsilon^{-}(i,k,t) \quad \forall i,k,t$$

$$-N(i,j) \cdot M \le f(i,j,t) \le N(i,j) \cdot M \ \forall i,j,t$$

$$\beta(i,j) \le \begin{cases} 0 \text{ if } i = j \\ M \text{ otherwise} \end{cases}$$

$$\beta(i,j) \ge \begin{cases} 0 \text{ if } i = j \\ -M \text{ otherwise} \end{cases}$$

$$\sum_{j=1}^{N_{TF}} N(i,j) \ge 1, \ \forall i$$

$$\sum_{i=1}^{N_{TF}} N(i,j) \ge 1, \ \forall j$$

$$\epsilon^{+}(i,k,t), \epsilon^{-}(i,k,t) \in \mathbb{R}^{+} \ \forall i,k,t$$

$$f(i,j,t), \beta(i,k) \in \mathbb{R} \ \forall i,j,k,t$$

$$i = 1,\ldots,N_{TF}, k = 1,\ldots,Nc, t = 1,\ldots,T$$

$$N(i,j) = \begin{cases} 1 \text{ if TF i regulates TF j} \\ 0 \text{ otherwise} \end{cases}$$

$$s(i,k) = \begin{cases} 1 \text{ if soluble factor k activates TF k} \\ \text{otherwise} \end{cases}$$

The variables β(i,k) indicate the level of activation of TF i in the presence of soluble signal j, Nc denotes the total number of simultaneous stimulation experiments. The introduction of this term was necessary because in the experimental design, there is no guarantee that each reporter plasmid will respond in the same way to an identical level of its stimulatory factor. For a single time point, there is a total of n + n^2 variables which need to be addressed.. The LCA allows us to compensate for this through the use of composite inputs in which multiple stimulatory factors are used at once thus allowing for the relatively easy introduction of another condition thus eliminating this problem. However, since β remains a constant, over the experimental time course, we actually have t * n^2 + n variables with t * n*(n + 1) equations thus making the system over-defined after the introduction of an additional condition. Formulation

Error! Reference source not found. concurrently reconciles the measurements based on k perturbation experiments. The L-1 norm is simulated through the use of appropriate positive slack variables. The prior information is hard-coded in the parameters N(i,j). The only provision at this point is that we assume that each TF has at least one regulator and that each factor regulates at least one member of the network.

This formulation requires the use of an *a priori* network architecture. While it is possible to obtain the information through the use of alternative experiment such as Chip-Chip experiment, or transcription factor prediction algorithm, it is also possible to derive the network architecture from our proposed bi-clustering algorithm which was presented previously. After the network architecture has been obtained via bi-clustering, it is reasonably easy specific the matrix N(i,j) to determine the underlying network architecture and solve for the dynamical response associated with the bi-clustering solution.

This is possible by converting the formulation given in **Error! Reference source not found.**, such that rather than utilize *a priori* information in N, N can be allowed to vary as a binary variable. Thus, rather than evaluating the equation for a given network architecture, we can instead allow the formulation to determine what the optimal network architecture ought to be. Therefore, rather than utilizing a set network as in the case of the previous formulation, we can have the formulation determine which network parameter is optimal. Furthermore, this formulation can be converted into a hybrid formulation allowing for both the incorporation of *a priori* information in the form of additional constraints, such that known connections are fixed, and unknown connections are allowed to vary in the formulation.

While the formulation can incorporate external information, we have not elected to do so. At this point, the focus is primarily upon the interaction of corticosteroids under direct stimulation

as well as dynamics of how corticosteroids impact the activity of other transcription factors. Secondly, our primary hypothesis is that through the reconstruction of the observed dynamics, it should be possible to hypothesize the existence of a given interaction with a high level of accuracy. Therefore, we wish to compare our unbiased results with what is currently known about the system.

Mathematical Programming

Again as in the case of our gene selection algorithm, the overall task is to solve an optimization problem. There exists multiple ways for obtaining a solution to the two sets of variables in which we need to find an optimal solution. For instance, it would be possible to use a modified least-squares approach to solve for the temporal dynamics for f(i,j,t), and either simulated annealing or genetic algorithms to solve for the binary variables. However, the drawback with all of these methods is that they are not guaranteed to converge to a global optimal. More problematic, the local solutions cannot be guaranteed to differ from the global solution by a set amount ε. But because the problem can be formulated with a closed form, it is possible for us to utilize a linear programming approach for solving the problem. The benefit of utilizing a linear programming approach is that a globally optimal solution can be found in a relatively efficient manner. Thus, the network which will be derived from the formulation is guaranteed to be optimal in terms of network architecture.

The disadvantage of utilizing a mathematical programming technique is that the computational complexity of the problem is not reduced. One of the difficulties with solving for the network architecture is that the problem is NP complete due to the combinatorial nature of the alternative network structures. What this optimizations approach does however is prune sub-optimal and infeasible solutions efficient. However, while it is able to reduce the number of

solutions that need to be evaluated fully, the problem still remains NP Complete. However, given our hypothesis that the emergent properties of biological systems are due to the NP Completeness, this computational complexity issue may not be one which should be ignored, or eliminated.

Evaluation of Dynamics

The overall hypothesis behind utilizing the RED is that the numerical response f(i,j,t) may provide insight as to the underlying processes which drive the observed changes in the activity of transcription factors. F(i,j,t) essentially represents how the various mechanisms transform the amount of active transcription factors into a signal which is then used to activate a secondary transcription factor. Treating the transcriptional network as a circuit analog, we can exploit the fact that many of the simple network architectures which we have obtained have well characterized step responses. Because the LCA utilizes a step input as the stimulatory profile for its soluble factors, we ought to be able to draw direct comparisons between the responses we see and the characteristic inputs. Previous work in electrical engineering has gone so far as to design automatic classifiers which categorize the step response based upon their circuit architecture[18, 140], we have elected to determine significant network architectures through visual inspection due to the significantly different responses of the network architectures.

We have elected to look for evidence of four types of dynamic interactions. The two which correspond to different network architectures [141] are feed forward [142] and feedback [143], whereas time lag, and tolerance mechanisms correspond to dynamic responses of the individual transcription factors. Identifying interaction motifs that eventually constitute the overall structure of a regulatory network is a very active research area and numerous methodologies have been developed to assess the emergence of local structures [144]. While there are other

methods for evaluating the statistical significance of each of the fits [145] we will be evaluating the possibility of developing specific network sub-structures by evaluating the dynamics of the interactions, f(i,j,t). The feed forward response represents the simplest response. In the feed forward interaction between transcription factors A→ B, strength of the up/down-regulation of B is dependent upon the activation of A and hence the activity of A's reporter, within a multiplicative factor. Time lagged dynamics can represent either intermediate transcription factors such as A→X→B in which X is an unknown factor, or events that that have a relatively slower rate limiting step such as the interaction between multiple sub-units. Feedback interactions emerge when the activation of transcription factor B, goes back and affects the activation of transcription factor A in addition to the standard feed forward response. The tolerance mechanism is a response which involves the loss of activation despite continued activation. In the LCA, there is a continuous infusion of the soluble signal, and therefore, this response should be quite evident. One of the possible mechanisms for this response is the loss of various receptors in the cytosol under continuous stimulation. **Figure 29** shows these basic interactions and the expected responses of the system. In addition to these simple models, the motifs can be combined for composite responses such as profiles that have both a time delay and tolerance effect.

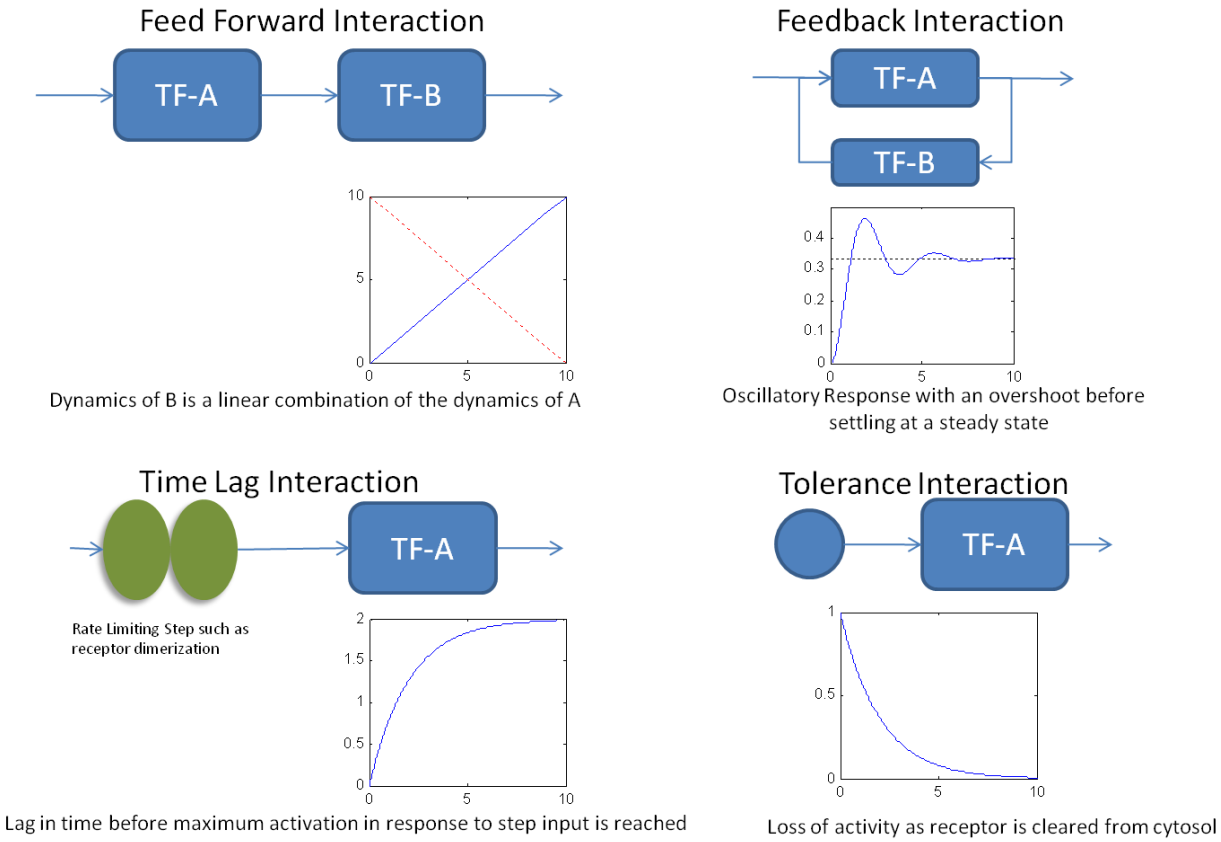


Figure 29: Expected network motifs and their expected responses. These interactions have a set response to a step input which is part of the experimental design. In these hypothetical interactions the x-axis represents time and the y-axis represents the interaction strengths

Trial Runs

The RED method was evaluated upon four different network architectures. It is hypothesized that by looking at several related network architectures, the ability of different network architectures may provide insights as to the importance of various network interactions. Specifically, we are looking for response which are similar over multiple network solutions to indicate the importance of a specific interaction, as well as coherent manners by which the network architecture can change. Central to this evaluation is the hypothesis that if the network architecture cannot be used to reconstruct the experimental data, we ought to see the loss of the motifs presented in the previous section. This essentially signals the loss of a possible mechanistic explanation for the experimental dynamics indicating that the network architecture is unable to replicate the experimental results. However, in the cases where the network architectures are different, and yet the dynamic motifs still exist it points to the possibility of "silent mutations" indicating the possibility of compensatory mechanisms that could be at play to compensate for the loss of the network architecture.

The four trials which we will run involve a fully connected network, in which all of the transcription factors are hypothesized to interact with every other transcription factor, the network generated via the bi-clustering formulism, a network which is optimal with respect to a given number of connections (6-36), and finally a hybrid network which consists of a freely optimized portion, and a portion of the network where the network architecture has been predefined. IN the case of the hybrid network architecture, the connections associated with AP-1 are fixed to 0, signaling that it will not be considered. This hybrid network represents a trial run that mixes the results of the bi-clustered network and freely optimized network. This was done so we could evaluate the effect of the bi-clustered network with respect to the other network architectures.

Results

Fully Connected Network

In **Figure 30** the profiles of the time varying weights are given. In this figure, there is evidence as to how each of the transcription factors interacts with others as well as themselves. Each row corresponds to the individual transcription factors whose activity we seek to reconstruct, and each column corresponds to the effect a specific transcription factor has upon the other factors within the system. For instance, the first row corresponds to the factors that affect the activity of NFkB. Likewise the first column corresponds to the effect NFkB has upon the other reporters.

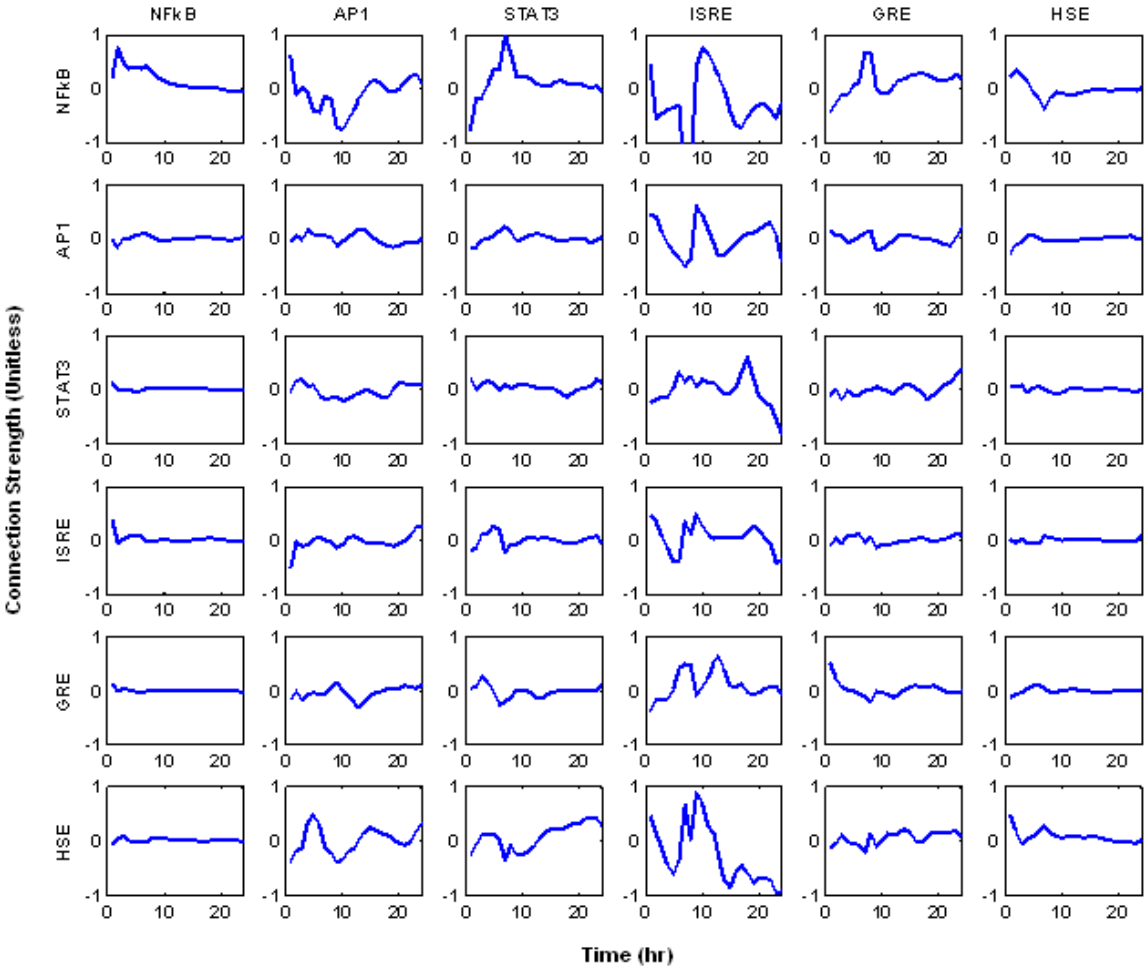


Figure 30: The dynamics of the interaction strengths calculated with a fully connected network. It is possible to see effects similar to those predicted via the motif patterns in Figure 29

In these figures, the weights that are negative represent a down-regulation effect whereas positive weights represent an up-regulation event i.e. an increase in the activity of one transcription factor decreases the activity of another and vice versa. From the profiles, we believe that evidence points to the fact that many of the dynamic processes are regulated by feedback control loops and therefore the simple notion that genes are only up or down-regulated may be too simplistic. For instance the stimulation of NFkB via GRE appears to be initially down-regulated, but also have a time period in which it is up-regulated after which it remains constant.

From the results, it may seem obvious which of the connections can be removed, i.e. those which show very low levels of activity. However, this may not always be the case. For instance, the STAT3→STAT3 interaction which corresponds to the stimulation of STAT3 by IL6 seems to be at a rather low level and can be removed. However, this connection needs to be included due to the design of the system. In which STAT3 is stimulated via its soluble factor. Therefore, it is not immediately obvious as to which connection should be removed. Such ambiguities therefore lead up to the next formulation in which the network is solved.

Freely Optimized Network

The freely optimized network solves the problem parametrically from 6 connections to 36 connections. The lower bound for the number of connections corresponds to the fact that each transcription factor needs to have some form of regulation, either via its soluble factor, or due to the effects of another transcription. The upper bound for the number of connections is the number of connections for a fully connected network. One of the problems with solving for the network in this manner is that it is difficult to tell *a priori* how many connections are needed. This then requires one to solve exhaustively for all possible number of connections. The trade-off

between complexity and the quality of fit is expressed as the pareto frontier. The pareto frontier indicated that

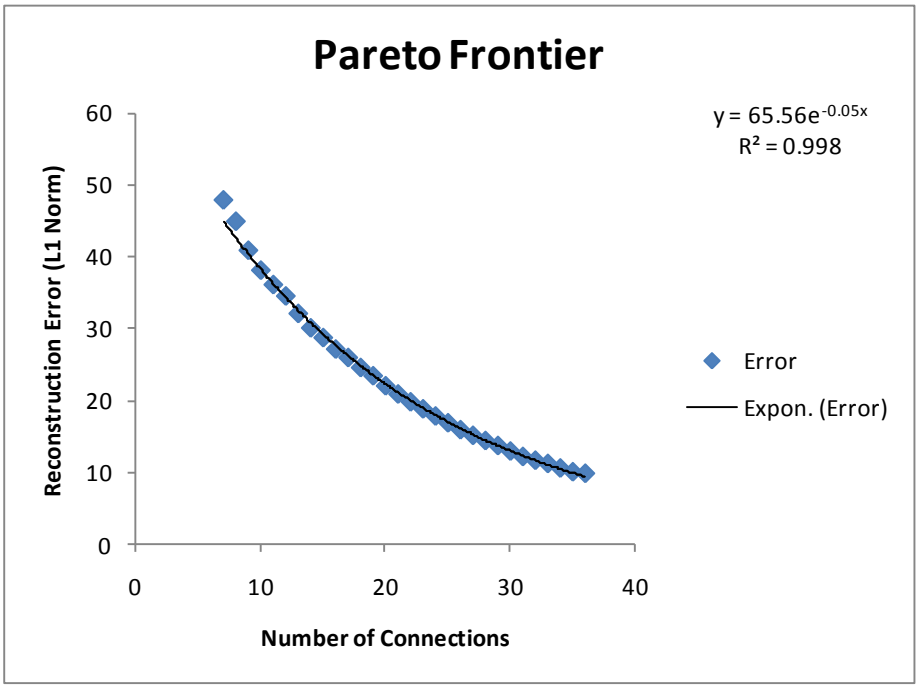


Figure 31: The pareto frontier. This plot is created in order to determine whether there existed a clear break in the objective function signaling the presence of redundant connections. However, we were unable to find this break.

one cannot obtain a better solution unless one increases the complexity of the problem being solved is given in **Figure 31.**

The pareto frontier for this system does not exhibit a “knee” feature which allows us to determine whether a sufficient number of connections have been obtained. It does however show an exponentially decaying response (R^2=.997), thus signifying that it obtains the relatively more important connections early rather than later. We hypothesize that the reason for this response is due to the small scale of the experimental data. Due to the fact that these transcription factors all related to inflammation, it is not surprisingly that all of the transcription factors may be part of a larger interconnected network. The small scale of the data means that many intermediate transcription factors are not present and therefore none of the links are truly redundant.

Though RED is unable to determine outright the number of connections present in the system, we hypothesized that it may still be able to give the relative importance of a given connection between two transcription factors. By solving the formulation from 6-36 connections, we expect the more important interactions to appear early and then to be conserved in solutions containing more connections. Therefore, is a set of interactions were present in a solution of size N, we would expect the great majority of the interactions would be present in solutions with more than N connection. Therefore, the most important interaction would be found first, and conserved throughout all the other solutions, the second most important interaction found second, etc. Plotting the number of times an interaction is present amongst the different solutions **Figure 32** exhibits this behavior. There is a smooth linear progression of importance,

whereas had the interactions been included at random, a connection would have been conserved an intermediate number of times (15) with a small amount of variability.

Given the structure of the pareto frontier, it is relatively difficult to determine the optimal number of connections. Therefore, the system was evaluated for 18 connections which correspond to the number of connections which have been identified via the bi-clustering technique. In **Figure 33**, the dynamics are shown. What is remarkable about the reconstructed dynamics is that for the links that are common between the networks most of the profiles seem similar in quality as in the fully connected network. This suggests that the reconstruction is reasonably robust not only in the way the individual links are incorporated, but also in the way the dynamics are obtained.

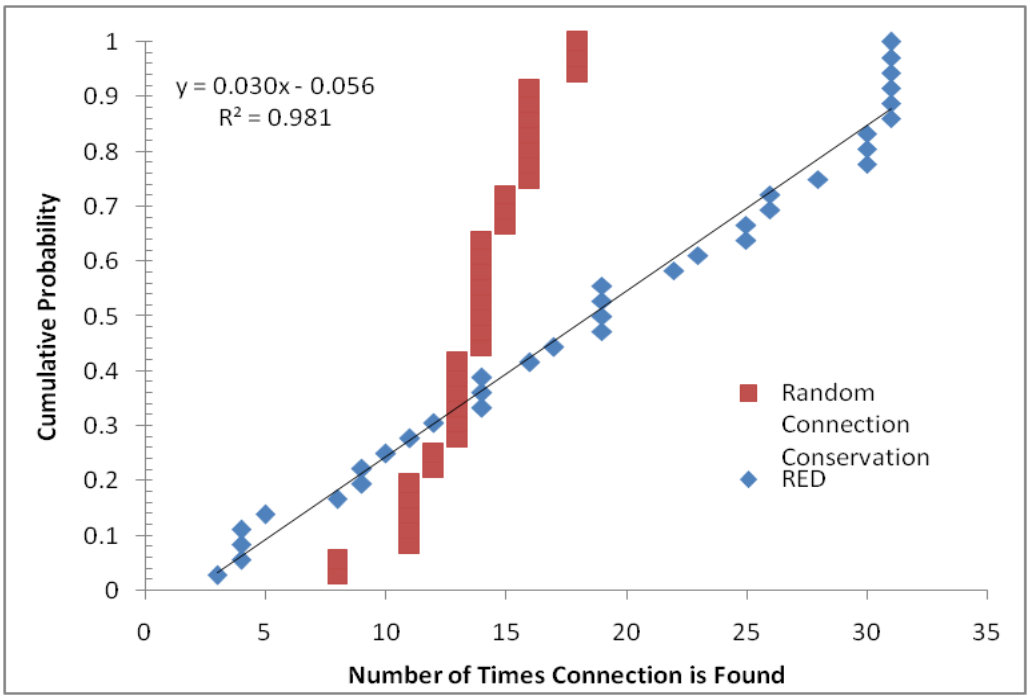


Figure 32: The number of times a link is conserved over the different solutions. Under RED there is a clear trend in the importance of links where as randomly assigned connects appear at a relatively consistent rate

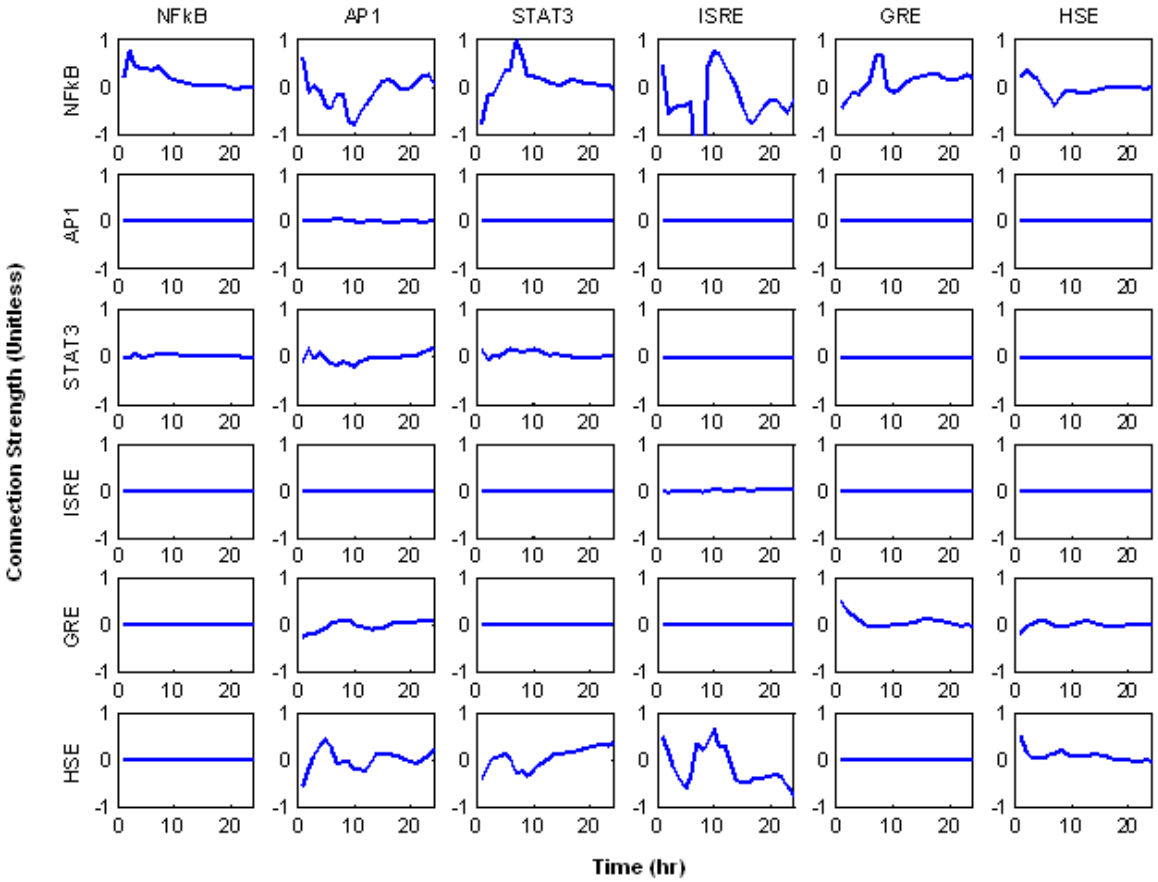


Figure 33: The dynamics associated with the freely optimized network utilizing 18 connections. The notable feature which we wanted to validate was the fact that the dynamics for A(i,i',t), are consistent over multiple solutions. The consistency of the dynamics suggests to us that our formulation is able to solve for some consistent underlying structure and response.

Bi-clustered Network

Rather than utilizing a fully connected network, it is possible to utilize bi-clustering to obtain the underlying network architecture, and observe how this predicted network architecture is able to replicate the experimental dynamics. From the bi-clustering result and the associated bipartite network **Figure 34**, it was found that while HSE did not have a specific activator under the experimental conditions, it showed significant co-expression and activation from a variety of different. The activation of the Heat Shock Element normally occurs in temperature above 35 degrees, and yet it was activated under the administrations of Dexamethasone, IL-6, and Interferon Gamma. The possible transduction of the HSE by Interferon Gamma has identified [146]. The activation by Dexamethasone has been previously identified but is weak and like the other results involving Dexamethasone, this may be more of an artifact off the poor data obtained via the administration of Dexamethasone. However perhaps as a reason for the poor results, the administration of Dexamethasone has been shown to either act as an antagonist for the binding of the heat shock element as well as increasing the production of the heat shock protein. Therefore, the poor results obtained from the LCA may be indicative of more complex behavior, for which all of the variables have not been adequately controlled.

Incorporating the *a priori* information which comes from the construction of the LCA, the directed graph given in **Figure 35** was obtained**.** In this figure, HSE does not have any direct stimulus because it was not stimulated by a given soluble factor. Secondly, IL1 was not found to stimulate any of the other factors in a manner coherent with the response of AP1, and therefore was not included in any of the bi-clusters. This is reflected in the fact that IL1-> AP1 appear to be disjoint from the rest of the network.

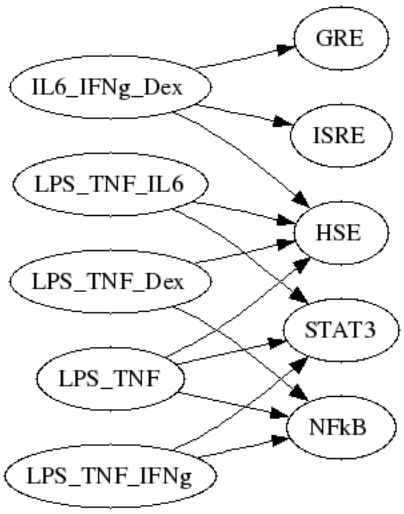


Figure 34: The bi-partite network associated with the bi-clustering solution

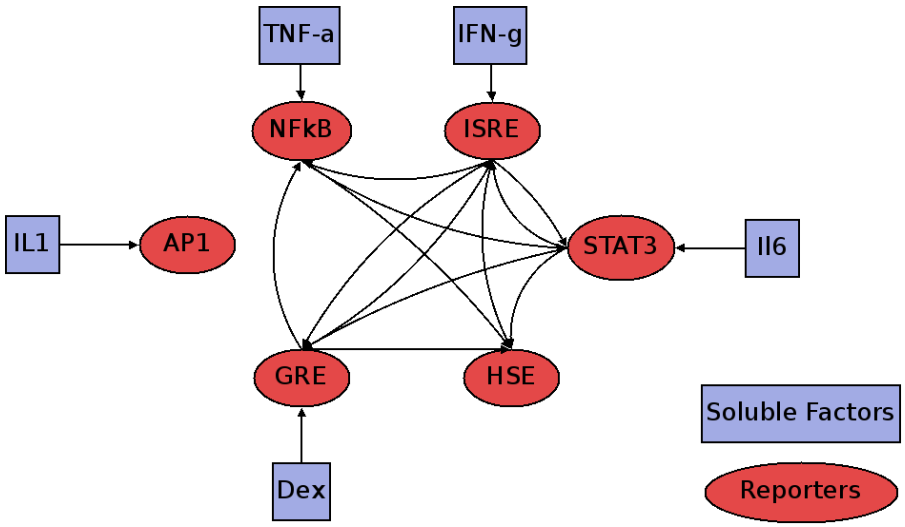


Figure 35: The directed graph that is equivalent to the bi-partite graph obtained via the bi-clustering result. HSE does not have a soluble factor associated with it due to the experimental design where it was not directly stimulated. IL1->AP1 was not included because it was not found to be present in any bi-cluster.

One of the concerns which we have with the results of both the bi-clustering as well as the network reconstruction is the effect of noisy data. One of the drawbacks of most clustering methods is that they oftentimes cluster all of the data without regard to data quality. Given the fact that our biclustering is highly dependent upon the initial clustering, any shortcomings due to noisy data would thereby be carried over to the generated network.

After the network had been generated, the resultant network was fed into the RED formulation to determine how well the network architecture can be used to replicate the underlying data. The primary reason as stated previously was that the use of *a priori* information could greatly reduce the computational complexity of the problem by reducing either the number of free binary variables or eliminating them outright. While the bi-clustering formulation is also an MILP formulation, the bi-clustering formulation scales better in terms of the number of binary variables needed for a given problem size having 2N binary variables as opposed to N^2+ N binary variables as in the case of the fully optimized network. Therefore, one of the questions is what the trade-off between runtime and reconstruction error is. Since the bi-clustered network itself consists of 18 connections, this result was compared with the fully optimized network with 18 connections.

Normally the simplest method for assessing the “correctness” of a network is to assess the error associated with the reconstruction. In the presented formulation this is indicated by the L1 norm, which is the sum of the positive and negative slacks. The overall range of possible errors ranges from above fifty to a minimum of 9.9. Since in the fully optimized network, we attempt to select network architectures with the lowest error, we have to determine whether the bi-clustered network represents a good trade-off between the ability to reconstruct the dynamics

and its decrease in run time. To evaluate this, we wanted to see whether on average randomly generated networks have a higher error associated with them than a bi-clustered network.

Generating 1000 random networks, we found that the mean error for the networks was 36.8632 with a standard deviation of 4.2, whereas the bi-clustered network corresponded to an error of 30.31. Therefore, while the bi-clustered network does not reconstruct the profiles as accurately as either the freely optimized network or the fully constructed network, we hypothesize that it does capture many salient features of these networks due to a reconstruction error which is significantly lower than that of a randomly generated network. Therefore, while the bi-clustered network does not yield an optimal reconstruction, it may function as an adequate approximation of the structures present within a given biological network. One of the advantages of utilizing bi-clustering to first determine the underlying structure is rather than the freely optimizing the network structure is the fact that bi-clustering coupled with the formulation in **Error! Reference source not found.** yields an operation that requires less binary variables which at a first approximation yields far lower runtimes. Therefore, the use of bi-clustering may be considered as a trade-off between accuracy and run time.

Because there are differences between the network architecture generated via RED and bi-clustering, we wish to determine whether these differences lead to changes in the dynamics of A(i,i',t) **Figure 36**, and whether the changes in dynamics can be explained due to mechanistic differences between the two solutions. For instance the NFkB/GRE interaction appears to have moved from a feedback dynamic to a standard feed forward interaction. The most glaring differences between the networks generated via RED and the ones generated via bi-clustering is fact that AP-1 no longer affects or is affected by the rest the network, and HSE does not have any outgoing connections. Given the qualitative differences between the network, the question

which arises is whether the change in dynamics is due to the loss of AP-1's affect upon the system or HSE's effect upon the system.

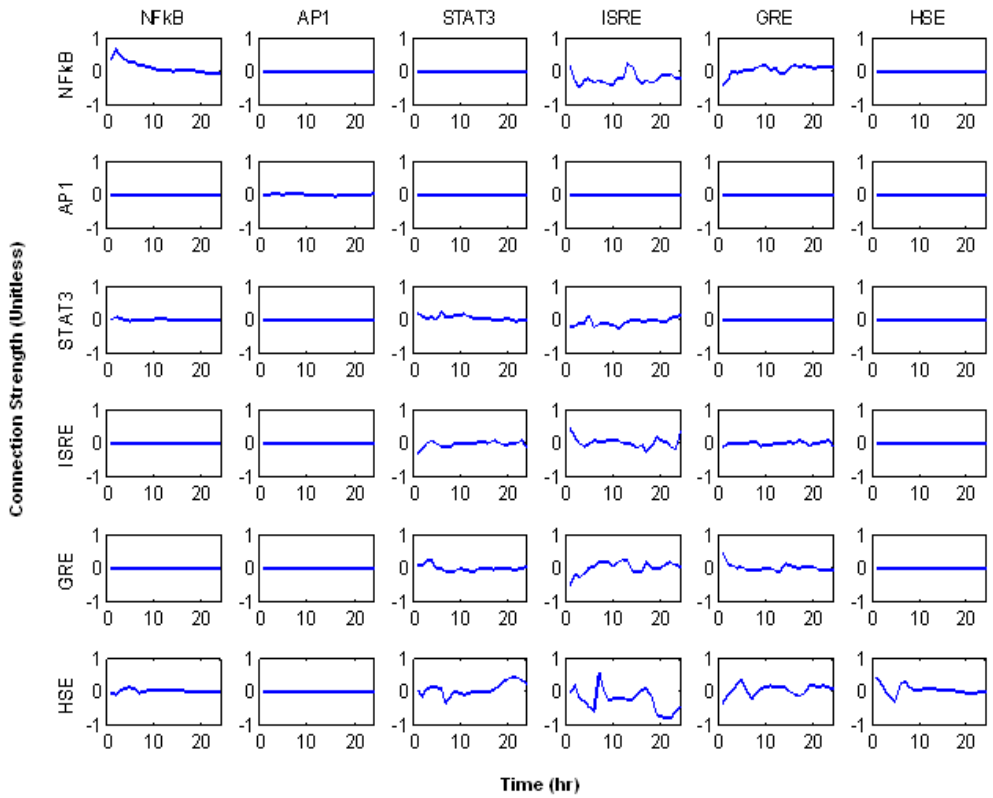


Figure 36: The reconstructed dynamics of the network obtained via bi-clustering. What is notable is the change of the response of NFkB to Dexamethasone stimulation (GRE) which turned from more of a direct interaction from a feedback interaction

Constrained Optimized Network

One of the things which we observed with the bi-clustered network was the fact that AP-1 was not incorporated into the network, and there were no outgoing links for HSE. Therefore, we manually remove the connections associated with AP-1 and then allow the optimizations framework to evaluate the presence of the rest of the network. This allows us to assess the role of HSE independently of AP-1. Doing so, we can see that the feedback dynamics associated with NFkB in response to GRE stimulation has returned as well as the response to ISRE. From this result it appears that the response of NFkB to both IFN-γ and Dexamethasone are in part affected by HSE. Moreover, without AP-1 affecting the dynamics of HSE, we see a large change in the dynamics of other transcription factors, thereby suggesting that AP-1 plays an important role in HSE activation.

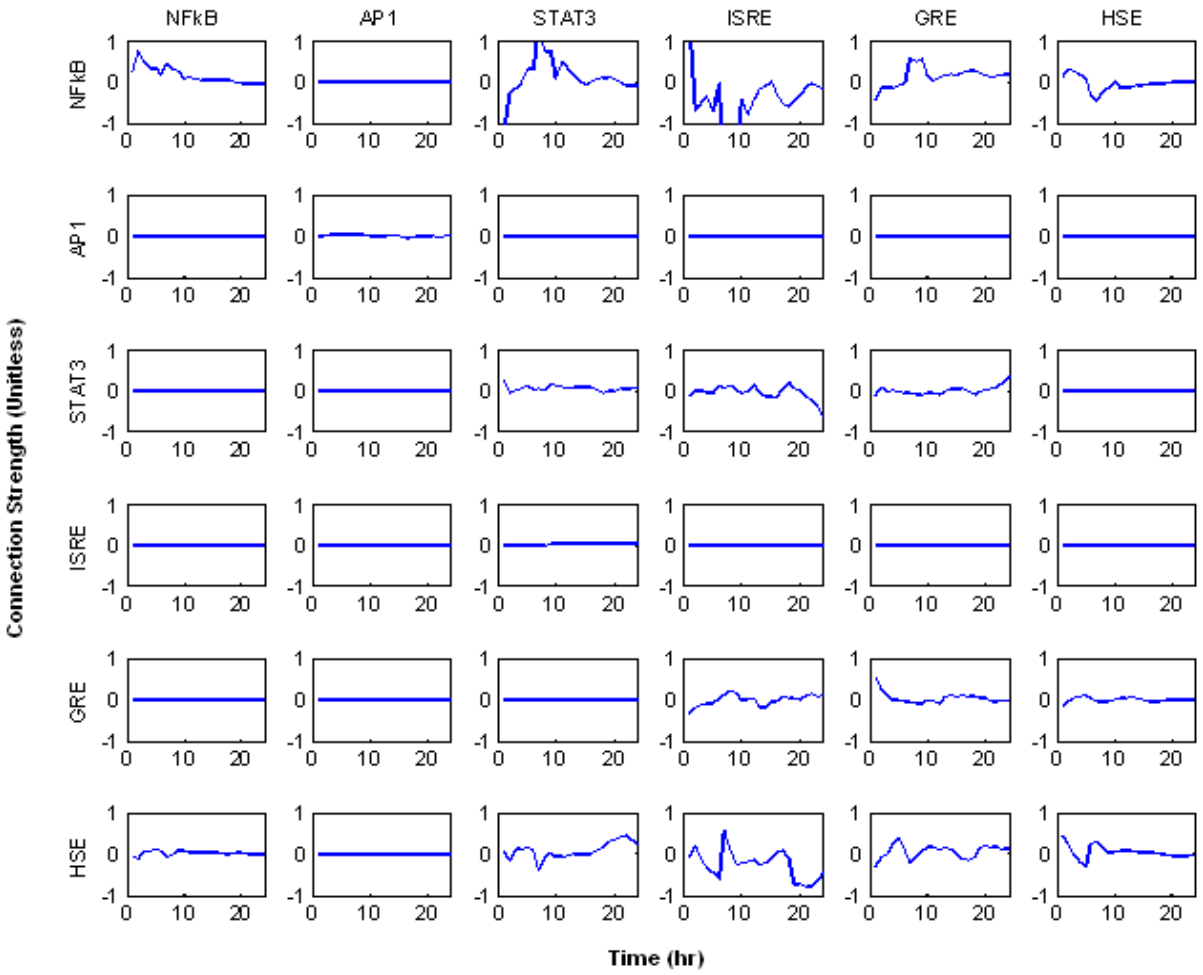


Figure 37: The response of the system when outgoing nodes from HSE were enabled, but the outgoing connections from AP1 were not. With the elimination of AP1, but the inclusion of HSE, we can observe many of the dynamics of NFkB's response to the other factors returning

Response to an administration of Corticosteroids

Of primary interest to us is the response of the corticosteroid receptor in response to activation by Dexamethasone. This dynamic is illustrated by the entry indicated as GRE-GRE in all of the Figures **30,33,36,37**, i.e. the stimulation of GRE by itself which denotes the effect of corticosteroid activation due to its stimulatory factor Dexamethasone. What is evident is the monotonically decaying dynamic of the system in response an infusion of corticosteroids. What is not evident however is the presence of a time lagged effect, or a sigmoidal response which would be indicative of some time-lagged interaction, or the presence of cooperative binding. With respect to the response motifs that have been presented in the previous section, the only nonlinear response which is evident under this dosing strategy is one of tolerance, in which the strength of the systemic response decreases over time.

Of secondary concern to us, is the response of NFkB to the administration of corticosteroid. Again the dynamics are present in Figure 5 and Figure 8 under the row labeled NFkB and the column labeled GRE. In this response, we can see evidence of an oscillatory behavior which may indicate the presence of a feedback mechanism. Specifically, the strength of Dexamethasone's activity upon NFkB is not mediated solely by the concentration of corticosteroids within the system, nor the loss of activity of the GRE receptor. One of the important observations which we make from this experiment and analysis is that the rapid loss of activity seen with inflammation related genes appears to occur faster than the decline and GRE activity and results independently of other tissues. More importantly, this deactivation of the inflammatory response occurs in an oscillatory manner suggesting the presence of a feedback mechanism mediating the response.

One interesting observation which was obtained from evaluating the different network architectures is that while the bi-clustered network exhibits a different response to corticosteroid than the other three networks, it switches from one dynamic motif to another. Thus, rather than having the classic feedback response that was exhibited in the other three solutions, it exhibited a standard tolerance response which was not unlike the response which was seen in the response of the GRE to a step input of corticosteroids. Thus, unlike the response of IFN-γ, it appears that this network architecture is still able to reconstruct the observed dynamics provided that a shift in the underlying mechanism occurs.

Discussion

One of the limitations associated with this implementation of the LCA, was our inability to determine the proper number of connections within the system. Given the plot of the pareto-frontier, an obvious cutoff does not appear to exist, where we can say that by increasing the complexity of the system, there is a negligible improvement in the reconstruction of the experimental data. Therefore, utilizing RED to solve for the network architecture as well as the network dynamics appears to suggest that a fully connected network is the most optimal network structure. While this result was disappointing, given the small scale of the experiment as well as the fact that all of the transcription factors were related to inflammation, this was not surprising. This indicates to us that there are a significant number of inflammatory transcription factors which have not been measured by our system.

However, in spite of our inability to determine the underlying network architecture, the use of RED to obtain the underlying network architecture shows an interesting result. It appears that as we increase the complexity of the solution in a parametric manner that the connections are conserved over multiple solutions. Thus, the majority of the connections associated with a

solution of complexity N, will be present in a solution of complexity N+1. This is important because it means that the formulation adds the connections in some sort of rational order. Because of the fact that these connections have been added in some rational order, we hypothesize that the resultant order may indicate some measure of biological significance. For instance, it was found that the response of NFkB to the other factors was more highly conserved than the activation of the other factors by NFkB. This is coupled with the fact that NFkB's activation of the other factors seem to take place at a low level. Thus, a reasonable hypothesis is that NFkB plays an important terminal role in the response of the system to the other inflammatory factors tested in the system.

By utilizing the different network architectures as well as the solve dynamics, we hypothesize that it may be possible to extract significant information as to the workings of the glucocorticosteroid upon the rest of the system. One of the important aspects of corticosteroid stimulation is the fact that under direct stimulation of corticosteroids, the signals obtained for processing were very noisy **Figure 12** as evidenced by the lack of repeatability in the measurements**.** The reason for this lack of signal fidelity is due to the fact that the majority of the inflammatory cytokines are down-regulated by corticosteroids. Working off a baseline fluorescence of the reporters being zero, the down-regulation of this signal means that the measurements are dominated by noise. To compensate for this, the experimental system also included composite stimulus represented by an infusion of all the inflammatory cytokines with the addition of Dexamethasone. This allows for a baseline fluorescence to be obtained and the effect of Dexamethasone to be de-convolved from the system. The ability for these interactions to be de-convolved is an important one because it shows that composite stimuli can be used successfully in the optimization framework and presents less of a problem in the network generation than the bi-clustering formulation.

The response of the system to Dexamethasone provides similar insights into the mechanism of corticosteroid activity. Unlike the response of NFkB to a step input of TNF-α, the response of the Dexamethasone upon its reporter GRE is a decreasing function indicative of a tolerance mechanism **Figures 29,32,35,36**. Therefore, the maximum effect of corticosteroids occurs early and there is no delay before a maximum is reached. This suggests that unlike NFkB, there is no rate limiting step between the binding of the corticosteroid to the glucocorticosteroid receptor (GR) and the activation of the transcription factor. This suggests that if there is a dimerization event is not rate-limited i.e. there is a sufficiently high concentration of endogenous GR present in the system which the rate of dimerization occurs fast enough where it is not detectable given the time resolution of our system.

Another interesting aspect of Dexamethasone stimulation is the response of the NFkB reporter in response to activated GRE. The currently accepted notion is that the activation of GRE down-regulated NFkB, thus damping the inflammatory response. However, one of the interesting aspects of this response is evident when the response over different solutions is compared. In the bi-clustering solution where the contributions from STAT3 and HSE were not included in the dynamics of NFkB, we see a clear shift from a feedback mechanism to that of a standard feed-forward response in which the strength of GRE's effect upon NFkB is directly related to the amount that GRE is stimulated. This suggests that unlike AP1, STAT3 and HSE play an important role in mediating the feedback mechanism that regulates the response of NFkB to corticosteroids. The presence of a feedback mechanism associated with the activation of the glucocorticosteroid receptor and the activation of the inflammatory cytokines is an important aspect that we will need to incorporate into our modeling.

However, in addition to the evaluations based upon the stimulation of the system by corticosteroids, a great deal more information may be extracted from this system. Of further interest to us may be the dynamics of how the other transcription factors interact with each other. Thus, we have included analysis of the other dynamics present in the system.

Predicted Result of NFkB Activation

The transcription factor interaction with the clearest activity profile is the activation of NFkB in response to TNF-α stimulation. This profile was present in all four solutions **Figure 29,32,35,36.** It is possible to see a clear lag in the activity of the transcription factor to the step response, in which it takes a non-trivial amount of time to reach a maximum, after which there is a return to baseline. This is indicative of a time lagged response coupled with a tolerance mechanism. This dynamic provides evidence of a rate limiting event in NFkB thus accounting for time lag before a maximum value is reached. This rate limiting step could be due to the time it takes for the subunits to be released from IkB[147] or via a rate limiting dimerization step. The return to baseline despite continuous infusion of the TNF-α signal illustrates that the NFkB shows a tolerance like response under prolonged administration of inflammatory cytokines[148].

One result which is unexpected is the fact that NFkB activation seems to have a low level of effect upon the other inflammatory cytokines, but seems to be significantly affected by the activity of the other inflammatory cytokines. This suggest that while TNF-α is an important mediator of inflammation, its reporter NFkB lies downstream in comparison to the other inflammatory transcription factors which were measured in this experiment. This result was consistent over the different solutions which leads us to the belief that this response is both highly robust, as well as the fact that the experiment yielded data of high quality for this particular transcription factor.

Predicted Result of AP1 Activation

The perturbation of the system through an administration of IL1 did not seem to have a large impact upon the AP-1 reporter in any of the solutions. This suggests that the reporter gene for AP-1 activation may need to be optimized. The level of up/down regulation of the AP-1 reporter in response to IL1 activation is low in contrast to the dynamics seen via the HSE, and NFkB reporters. This was evident in the fully connected and freely optimized networks **Figures 29,32**, and as such the effect of the soluble factor IL1 is evident within the system, though not through its individual reporter.

In the solutions of the fully connected and freely optimized networks given in **Figure 29,32** there appears to be a feedback mechanism associated with AP1 and HSE, with an oscillatory behavior in the weights. We see that in the freely optimized network, there is still a great deal of commonality in the response of A(i,i',t) between the two cases despite the removal of the effects of GRE and NFkB upon the system. With the removal of AP1 however, we see that there exists a major change in the dynamics of the rest of the system **Figure 35,36**. Computationally, this means that the loss of AP1 as a connection, requires significant alterations to the dynamics of other transcription factors in order to fit the data. We hypothesize that this effect is due to HSE repressing the synthesis of IL1 which is the activator of AP-1[149], and that IL1 affects the phosphorylation of various heat shock proteins[150]. The combination of these two factors suggests the existence of a cycle and therefore the need for a feedback interaction between the two elements.

It may be tempting to suggest that the effect of AP1 upon the rest of the system can predicted merely the magnitude of its interaction. However, this is not necessarily the case as seen in the interactions of the other transcription factors with NFkB. What we see is that even with the

removal of AP1 **Figure 36** from the solution, the interaction dynamics of the other transcription factors with NFkB are still reasonably consistent and that it requires the removal of both HSE and AP1 before there exists a significant change.

Predicted Result of STAT3 Activation

Similar to the results obtained with the IL1 stimulation, the IL6 stimulation did not appear to have a large effect upon the induction of its reporter. This dynamic was again present in all of the solutions that were obtained. This evidence suggests that the sequence of the reporter could perhaps be better designed. Specifically, while these reporters are able to show qualitative changes, they may be optimized to show greater fold change when activated. In spite of the low fold change in STAT3 reporter activation, there was a significant alteration in the activity of the NFkB reporter by IL6. This was present in three of the four solutions **Figures 29,32,36** being present in the solutions where the network was determined via the MILP formulation and absent when utilizing the bi-clustered network. This activation appears to have a feedback-type dynamic for NFkB. In the literature, it has been reported that IL6 is induced by TNF-α, an activator of NFkB[151] as well evidence that IL6 down-regulates the activity of NFkB[152]. This combination of effects points to the existence of the feedback mechanism as suggested via the reconstructed dynamics. The other reporters in response to IL6 stimulation have inconsistent results and connectivity. Given that the connection strengths of STAT3 to the other transcription factors is low, this suggests that perhaps the connections may not actually exist, or that IL6 represents a relatively non-specific inducer of inflammation and that while it affects many system, its individual contribution to the dynamics of the measured reporters is reasonably low.

Predicted Result of ISRE Activation

While many of the transcription factor interactions seem to have dynamics which are similar to those predicted via the network motifs in **Figure 29**, the responses for IFN-γ stimulation do not. This may be due to the highly connected nature of IFN-γ due to its central role in the JAK-STAT pathway[153]. Aside from the interconnectedness of the IFN-γ, the networks generated via the full optimization, and the bi-clustering both appear to reflect the fact that ISRE is consistently more connected than any of the other elements. Due to small scale of the system, the effect of IFN-γ on the other factors may be in reality mediated through several intermediates. Without these intermediates, the effect of IFN-γ upon the system represents the combination of the effects of these different intermediates, thereby obscuring the direct effect that IFN-γ has upon the system. However, the pseudo-oscillatory behavior may be indicative of a significant amount of feedback that underlies an organism's response to IFN-γ, and may be due to factors which have not been previously identified.

Conclusion

The primary pieces of information which we wished to extract from the Living Cell Array were the hypotheses that the tolerance effect of corticosteroids appeared to have been mediated by receptor saturation, and not through the degradation of the receptor itself. Thus, the tolerance mechanism associated with its anti-inflammatory effects is independent of the concentration of activated receptor within the system. Secondly, it appears that the loss of the anti-inflammatory effects of corticosteroids are mediated specifically by a feedback mechanism which attenuates the anti-inflammatory effect of corticosteroids, as evidenced by the oscillations present in the response of NFkB to an activated corticosteroid receptor.

Furthermore, we hypothesize that such a feedback mechanism may be mediated by either a secondary transcription factor, or a protein which is regulated by this transcription factor. With the loss of AP-1 and HSE, we see the loss of the oscillations in the response. Thus, a secondary hypothesis is that the feedback mechanism may be affected by one or both of these transcription factors. With the removal of these transcription factors, the effect of corticosteroids on NFkB appears to follow the same saturation kinetics as response of the GRE reporter. These pieces of information then many allow us to create a more comprehensive model of corticosteroid activity.

Aside from the specific information about the dynamics of the corticosteroid receptor, the formulation appears to be able to distinguish the dynamics associated with various mechanistic of transcription factor activity such as receptor dimerization. Thus, this system exemplifies the goal of systems biology. With this system, we can run a relatively standard experiment, combine the experimental data with an analysis technique which is implicitly takes the interactions of the different systems into account, and finally obtain a hypothesis as to either the mechanism behind the activation of a specific transcription factor as well as the possible architecture which governs the entire response of an organism. This hypothesis can later be tested more thoroughly in more precise and targeted experiments.

One of the logical progressions of this method is the expansion of the Living Cell Array to incorporate more transcription factors. However, in addition to the experimental scale up of the device, further work also needs to be done in optimizing the formulation such that larger systems can be tackled. While the number of binary variables in the bi-clustering algorithm scales linearly with the number of factors added into the system, the RED method scales in polynomially, which quickly makes the problem intractable. This may be solvable through a more intelligent use of constraints in the formulation, or through the incorporation of *a priori*

information. The incorporation of *a priori* information is attractive to us, because it suggests the ability to link the experimental work and the analysis in an iterative loop which allows for constant refinement of the predicted architecture and dynamics.

A Hypothetical Model for CEquation Chapter (Next) Section 1 orticosteroid activity

Because this dissertation represents a Systems Biology Approach for assessing corticosteroid activity, the final step is the proposal of a new hypothesis that can be used to guide further experiments. Thus, the final step of this dissertation is the proposal of a new model of corticosteroid activity which one can then test utilizing either different input stimulus, or by testing for certain important features within the data. Therefore, we will not only be proposing a model in the final chapter of this dissertation, but also suggesting different experiments which can be run to validate such a model.

One of the issues with the current model for corticosteroid activity was that it was developed under an acute administration of corticosteroid administration. Utilizing the results from the SLINGSHOTS algorithm as well as a basic mass transfer model, the acute response can be explained via a simple mass-action model which suggests that the primary effect of the drug is established through the local dug concentration in different compartments. Therefore, while there does exist significant nonlinear effects hypothesized by Almon et al, such as tolerance, this dataset does not exhibit the necessary features to resolve them. Under the chronic administration of corticosteroids, we see two characteristic responses, the first response which is similar to the profiles obtained via the acute administration, and a second response which showed an initial lag, and then a sustained up-regulation of the gene expression profile.

Paradoxically, the gene expression profiles which appear to be common between the acute and chronic administrations of corticosteroids are dissimilar mechanistically. Under a chronic infusion of corticosteroids, the expected response of a mass transfer model was given in **Figure 16**, which is a sustained up-regulation of activity due to the continued presence of the activated

transcription factor within the system, with the inclusion of a discrete lag term which we do not model in our simple mass transfer system. Utilizing the more complex fifth generation model, does not yield qualitatively better results. Because in both models, the primary driving force is the amount of activated corticosteroid within the system, and in both models, this is directly related to the amount of drug within the system.

In an updated mechanism by Almon et al., they proposed the fact that corticosteroids destabilize mRNA, and thus, a constant infusion of corticosteroids will degrade the mRNA associated with inflammation, and not those associated with metabolism, thus explaining the two responses. While this mechanism is a reasonable hypothesis, it does not explain the salient time lag feature in the gene expression profiles from the extracted chronic dosing of corticosteroids such as those seen in Cluster 1 of **Figure 14**. Furthermore, in the acute case, it was found that the different gene expression profiles had two sets of time constants, one of which had a relatively fast activation, and a second profile which had a similar activation profile, but one which took place at a later time. Because of these two observations, we believe that the effect of corticosteroids upon the liver is a two step sequential process, rather than the updated model proposed by Almon et al, which treated the responses as two separate profiles, with the sustained response being similar to our simple mass transfer model, and the profiles exhibiting a loss of activity after prolonged infusion being due to a secondary factor which affects mRNA degradation rates.[154]

Our hypothesis for a two step process opens three possibilities:

1. One of the initial fast response genes code for another corticosteroid sensitive transcription factor which shows a different characteristics than the GR receptor

2. The corticosteroid receptor itself is modified via the fast response into a second active form.

3. There exists a significant interplay between different tissue types which leads to a secondary activation of metabolic genes.

Hypothesis 1 is based upon the existence of a second corticosteroid responsive protein which has different characteristics than the known corticosteroid receptor. This possibility has been touched upon in our transcription factor evaluation suggesting that AP2-α may be such a factor and the work by which showed that the anti-inflammatory effects of corticosteroids could be obtained even when the corticosteroid receptor has been re-engineered such that the sub-units could not dimerize and therefore activate[155]. In addition to the identity of a secondary corticosteroid sensitive protein, it also requires a relatively rapid degradation of the original corticosteroid receptor under a chronic infusion of corticosteroids.

Hypothesis 2 suggests that the corticosteroid receptor itself has two distinct active forms. The first active form is a drug bound monomer, and the second active form is a drug bound dimer, each of which recognizes different upstream cis-binding sites. Thus, under chronic stimulation the initial “fast” response is activated by one form of the corticosteroid receptor, codes for a protein that converts this corticosteroid receptor into a second form, which then goes on to activate the secondary sustained response.

Hypothesis 3 suggests that as corticosteroids act upon the different tissues in an organism, changes in the levels of circulating metabolites activate a second set of genes. Thus, the responses seen under the chronic case are not necessarily sensitive to corticosteroids, but are sensitive to the level of metabolites found within circulation such as the increase in circulation glucose. Thus would suggest a second regulatory site involved in the response to corticosteroids.

Utilizing the LCA, we hypothesize that one may be able to identify the most probable scenario based upon the process of elimination. For instance, if one were unable to see the rapid degradation of the corticosteroid receptor in the LCA, then hypothesis 1 would be discounted. Hypothesis 2 would be discounted if the corticosteroid receptor showed dynamics which were more complex than simple degradation of the receptor, or saturation. Hypothesis 3 could be discounted if the observed responses of the system could be replicated within the confines of the LCA, which does not incorporate different cell types and focuses only upon the hepatic cells.

From the results of the LCA, it appears that Hypothesis 1 may not be the most likely explanation. Specifically looking at the activation of the glucocorticosteroid receptor by Dexamethasone, we see a slow decline in the effect of Dexamethasone. This loss of effect can be attributed either to the decreasing effectiveness of additional amounts of corticosteroids due to receptor saturation, or due to the loss of the corticosteroid receptor itself due to the degradation. However, in either case, the decline in the effect of corticosteroids occurs relatively slowly, and thus cannot explain the rapid inactivation present under the chronic administration of corticosteroids. Secondly, the oscillatory feedback behavior of the NFkB transcription factor seems to occur in a manner that is not completely dependent upon the level of corticosteroids within the system. What the results of the LCA analysis shows is that the while NFkB is originally down-regulated by the corticosteroid receptor, there exists some feedback mechanism which controls the down-regulation of NFkB, and that this feedback mechanism acts faster than the decline in the glucocorticosteroid receptor activity. Thus, if a gene related to inflammation were related to both the concentration of corticosteroid receptor, and the presence of some other transcription factor, we would expect to see a response which was sustained longer due to the time duration of the glucocorticosteroid receptor.

The second hypothesis appears to be a variation of the first hypothesis. Rather than the transcription of a new protein, we have instead the conversion of the receptor from one active state to another. However, the primary difference between these two hypotheses is that given the structure of the regulatory regions built into the reporter plasmid that both forms are able to activate and bind to the promoter region, thus preventing the dynamics of the glucocorticosteroid receptor in the LCA from changing much. Furthermore, the existence of a dimeric active form, also allows us to rationalize the “time lag” response obtained from the gene expression data under chronic stimulation where there exists a small lag before the response increases to a new steady state. Furthermore, it incorporates an explicit model for the deactivation of the glucocorticosteroid receptor as a transcription factor, something which could not have been accounted for in the first hypothesis unless we hypothesize the existence of a third protein.

The final hypothesis was discounted because the sustained effect of dexamethasone on the activation of the glucocorticosteroid receptor as well as the shorter term loss of anti-inflammatory ability was seen in the LCA without requiring the use of other cell types. Therefore, while there may be an important role of other cells in the form of paracrine or endocrine signaling, the primary features which we are attempting to reconstruct are present when looking only at a single cell type.

Hypothesis: The glucocorticosteroid receptor has 2 active states

Given the results of our analysis of both the temporal gene expression data and that of the Living Cell Array, we have proposed the hypothesis that the corticosteroids act primarily through the corticosteroid receptor, and that this receptor has 2 active forms, the first form is a monomeric form and the second form is a dimeric form. More specifically, we hypothesize that

the monomeric form is the one which primarily mediates the anti-inflammatory and immuno-suppressive effects of corticosteroids whereas the dimeric form is the one which primarily regulates the metabolic response of corticosteroids.

The response of the Living Cell Array, particularly response of the glucocorticosteroid receptor to a sustained input of dexamethasone showed a monotonically decreasing function, where the maximum effect of corticosteroids was present during the initiation of the drug stimuli. This is in contrast to the response of NFkB, where the point of maximal drug activity occurs at an intermediate time point. From the motifs which we had previously indicated in ***Figure 30***, NfkB showed evidence of a time-lagged effect which could have been explained via the dimerization event of the NfkB subunits. However, such an event was not seen under the response of the glucocorticosteroid receptor under the administration of dexamethasone, which showed a smooth monotonically reduction in activity.

This suggests to us that there exists a state of the corticosteroid receptor which is active in lieu of dimerization. While it is possible that the time-lagged effect was not seen under the LCA data analysis due to the insufficient sampling rate, having an active monomeric form of corticosteroids allows us to resolve several paradoxical observations about corticosteroids in general.

For instance, it has been suggested that the anti-inflammatory effects of corticosteroids may not be mediated via transcriptional regulation. In a paper by Reichardt et al.[155], a mutant strain of mice was created such that the corticosteroid receptor could not dimerize. It was found that there was no loss in corticosteroid activity in these mutant rats. The conclusion drawn by Reichardt et al., was that corticosteroids do not inhibit inflammation transcriptionally. This interpretation of the results however does not take into account the natural metabolic effects

associated with corticosteroids. Had the corticosteroid receptor been truly non-active in these organisms, there ought to have been significant alterations to the metabolic response of the organism which should have manifested itself as phenomena such as hypoglycemia, loss of body fat percentage, and hypersensitivty to insulin. However, these phenomena had not been observed in the mutant rats. Therefore, an alternative hypothesis is that corticosteroids may be active in their monomeric form.

Furthermore, it was found that the corticosteroid receptor, even without dimerization, is able to bind to a corticosteroid half-site[156]. This was leveraged in the creation of ELISA assays for the corticosteroid receptor. In these assay, only the canonical half-site binding domain was used to bind corticosteroids. This at the very least suggests that whether the corticosteroid receptor itself dimerizes or not, it is at least able to bind to sequences expressing only one half-site.

The final paradoxical information arises from the documented position weight matrix associated with the glucocorticosteroid receptor. Given the current hypothesis that the corticosteroid receptor forms a homodimer, it would be expected that the canonical corticosteroid half-site of TGTTCT be consistent between both of the half sites, either directly or palindromic as shown in **Table 4**.

TCTTGTnnnTCTTGT
TCTTGTnnnTGTTCT
TCTTGTnnnAGAACA
TCTTGTnnnACAAGA

Table 4: Possible expected motifs if only the dimeric form were active

However, the currently accepted position weight matrix does not illustrate this dynamic, but is instead listed on TRANSFAC[69] as GGTACAANNTGTYCT, with which is not palindromic. Furthermore, one of the half sites is highly specific, whereas point mutations in the other half site is appears to have little effect upon binding[157]. Various possibilities exist such as the hypothesis that the second binding site changes based upon whether the corticosteroid receptor itself activates or represses the gene in question. However, while this would point to a loss in specificity, it does not explain the complete lack of specificity in this region. We do not see two specific possibilities manifest in the inconsistent region, but rather a random distribution of bases. Again, one method for rationalizing these observations is that there exist two forms of the active corticosteroid receptor.

Utilizing this hypothesis, one can rationalize the inconsistencies in the position weight matrix. Because one of the forms of the corticosteroid receptor represents the binding site of a monomer, taking a window size associated with the dimeric form will essentially yield a highly conserved region as well as a highly variable region which is similar to the properties observed in the position weight matrix of the glucocorticosteroid receptor.

Model Building

Having established our hypothesis that corticosteroids may exist in two different active forms, we can create a new model of corticosteroid activity. The primary feature of this model is the sequential activation of the different forms of the corticosteroid receptor. Therefore, the first form of corticosteroid receptor is activated, and will go on and either up/down-regulate the responses associated with inflammation and the immune system. The conversion of this first

active form into its second active form, will deactivate the receptor with respect to the first set of genes.

This leads to the following model schematic given in **Figure 38**. In this model, we hypothesize that the monomeric form represents the first activated state of the corticosteroid receptor, whereas the dimeric form represents the second activated state. This is because from the LCA data, the response of the system appeared to be driven primarily by the monomeric form. Therefore, the initial fast response appears to be primarily due to the monomeric form. At a later point in time, the dimeric form dominates, and should show smooth degradation/saturation kinetics. Given the hypothesis from the Living Cell Array, this initial response is mediated by a feedback mechanism. Therefore, one of the genes which are activated by the initial form of corticosteroids, will be part of the feedback mechanism which deactivates the receptor with respect to the initial set of activated gene. This thus, allows us to rationalize the dynamics observed with the NFkB transcription factor. The model consists of two linked receptor mediated dynamics. The initial response produces a protein which will convert the initial monomeric form of the receptor into a dimeric form. Because the genes which are sensitive to the monomeric form do not have the additional binding site associated with the dimeric form, transcriptional regulation of these genes decreases due to the inability to bind. This dimeric form then regulates the second wave response.

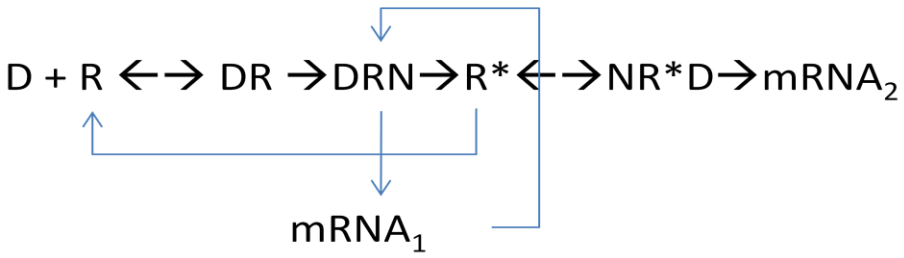


Figure 38: Our proposed model of corticosteroid activity. The primary features of this model are the feedback look which takes the protein produced via $mRNA_1$ which provides a feedback interaction, and the two active states of the glucocorticosteroid receptor D, and NR*

The set of differential equations associated with this model is given in (7.1)**.** One aspect of the model which we will not be explicitly modeling is the production and degradation of the corticosteroid receptor. While this was an important feature in the previous fifth generation model, we feel that there is not sufficient data to accurately model the temporal dynamics of corticosteroid receptor production and degradation. In the current dataset presented by Almon et al., the only piece of data that is available is the amount of free receptor. This is problematic because the amount of free receptor is dependent first upon the total amount of receptor present in the system, but is also dependent upon the amount of drug within the system. Because of the limited dataset, we cannot de-convolve these two processes. While undoubtedly the receptor dynamics will play a role in the overall response of the system, we hypothesize that the primary determinant of the response will be mediated via drug/receptor trafficking. We are not saying that the receptor production/degradation is not an important aspect of the system, but that under these two datasets, evidence of their impact may not be visible. This argument is similar to the argument presented in Chapter 4, in which models derived from the acute administration of corticosteroids could not be used to predict the response to chronic administration of corticosteroids, because not all of the nonlinear elements were evident in the response.

$$
\begin{aligned}
&D = c_1 * e^{(-\lambda_1 t)} + c_2 * e^{(-\lambda_2 * t)}; \text{Drug Concentration under acute dosing} \\
&D = 1; \text{Drug Concentration under Chronic dosing} \\
&\frac{d[R]}{dt} = -k_1[D][R] + k_2[DR] + k_{11}[R^*] \\
&\frac{d[DR]}{dt} = k_1[D][R] - k_2[DR] - k_3[DR] \\
&\frac{d[DRN]}{dt} = k_3[DR] - k_4[DRN][mRNA_1] \\
&\frac{d[mRNA_1]}{dt} = k_5[DRN] - k_6[mRNA_1] \\
&\frac{d[R^*]}{dt} = k_4[DRN][mRNA_1] - k_7[D]^2[R^*]^2 + k_8[DR^*] - k_{11}[R^*] \\
&\frac{d[DR^*]}{dt} = k_7[D]^2[R^*]^2 - k_8[DR^*] \\
&\frac{d[mRNA_2]}{dt} = k_9[DR^*] - k_{10}[mRNA_2]
\end{aligned}
\tag{7.1}
$$

In this model, D represents the concentration of the drug, R, represents the initial form of the receptor, DR, represents the drug receptor complex, DRN represents the drug receptor complex translocated into the nucleus, R*, represents the second active form, and likewise DR* represents the second active form complexed with the drug.

Model Fitting

Having established a hypothetical model, the question is whether, such a model can be used to replicate the dynamics found in both the acute and chronic administrations of corticosteroid. Working off of the elementary responses rather than the responses of individual genes, it is important to determine whether the qualitative responses can be fitted. From the results of the SLINGSHOTS selection, we have separated the motifs into two primary groups. In the acute case the clusters correspond to an early and late response **Fig.** The first cluster corresponds to the early response and the second cluster corresponds to the late response. In the chronic case, the two groups are divided into the responses which return to baseline, and those that do not. For

the sake of fitting, we select the groups that have the most genes as the representatives of all the other clusters.

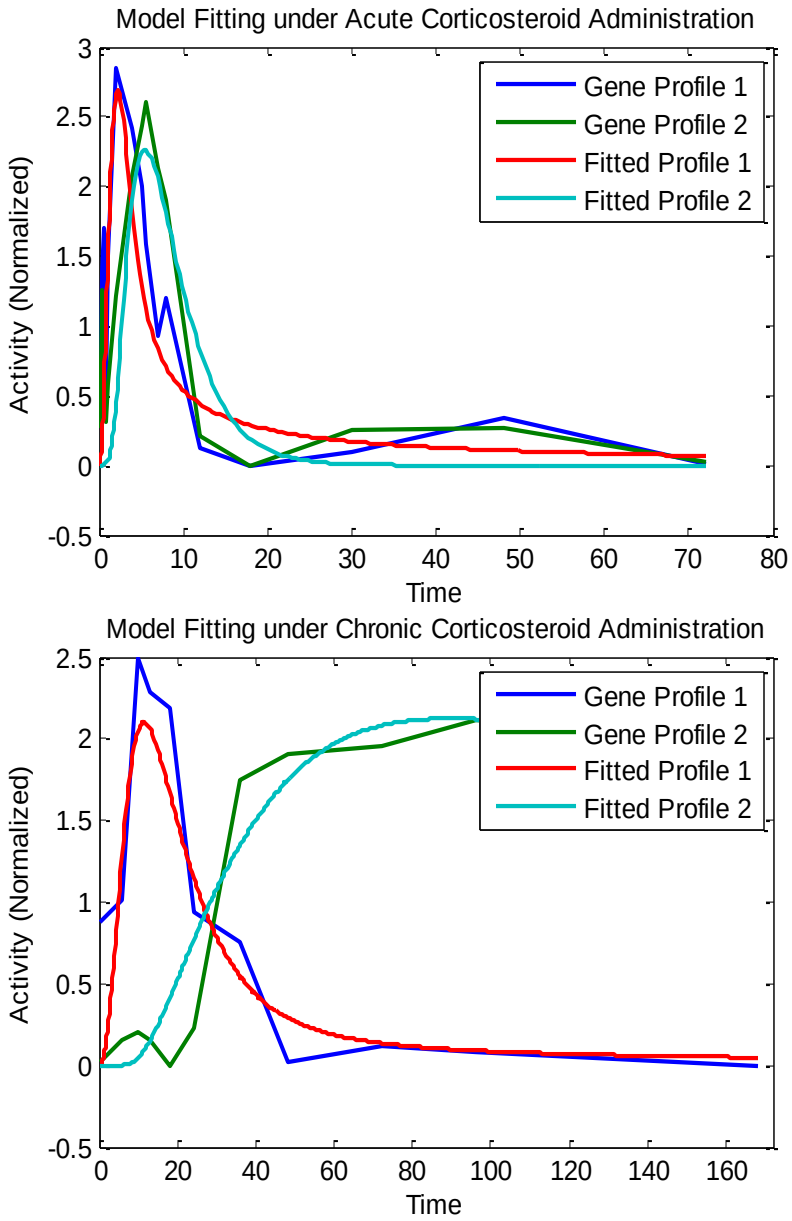


Figure 39: The profiles which were selected for fitting by our new model. The acute case consisted of the two up-regulated profiles. These profiles differ in the time constants associated with their time to maximum and return. The chronic case consisted of a profile which exhibited tolerance (loss of activity), and one which did not

One of the critical aspects of this model which we would like to stress is the fact that the underlying model is the same, and the only difference between the two models is the input into the system and the coefficients which are used to fit the model. The inputs of the system are defined by the drug concentration. In the acute case, we will be utilizing the bi-exponential function which was fitted to the drug concentrations. This function was explicitly defined as in Eqn. In the case of the step function, the drug concentration was assumed to be constant step function with a value of .1. The functional form of the drug concentration in circulation was defined with a dosing of 50 mg/kg whereas in the infusion experiment the drug concentration was defined with a dosing of .1 mg/kg, and therefore the amount of drug in circulation was scaled accordingly.

The data which we will be fitting is the elementary responses which we have selected via the SLINGSHOTS algorithm. Due to the fact that the SLINGHOSTS algorithm yielded many similar responses, we have selected two representative responses to conduct the fitting. In the acute cases, we have chosen clusters one and two, due to their different time constants. Of primary concern in this model is whether receptor trafficking can be used to explain most of the dynamics. Because Cluster 3 has a similar time constant to Cluster 2, it is treated as representing the same response except that it represents a repression event due to the binding of the glucocorticosteroid receptor rather than an activation event. In the chronic case, we have chosen as representative profiles Clusters 1 and 6. Cluster 1 represents the response which reaches a sustained maximum in response to continued drug administration whereas Cluster 6 represents the profiles that show tolerance i.e. loss of effect despite continued drug administration. In both these cases, the elementary responses were reduced to yield two classes of responses that were qualitatively similar. Thus after, we have validated the ability of the model to fit these curves, the model can be expanded to fit the response of all extracted clusters.

As a prelude to the fitting operation, the extracted motifs have been interpolated to have 1 hour intervals. Due to the uneven sampling of the experimental dataset, if no interpolation was done, then the early time points would contribute more towards the objective function. In this manner, the objective function weighted equally over the time course of the experiment. Due to this interpolation operation, the acute dataset has 72 time points which must be fitted and the chronic dataset has 168 time points which must be fitted.

In this model there are 12 parameters that must be fitted. The model fitting operation was carried out via the matlab command fminsearch. At this point, we wish to determine whether there exists a set of coefficients which will allow both of the observed responses to be fitted. At this point we are not concerned so much about the value of the different model coefficient, but only that the model can be used to reconstruct the observed dynamics.

In Figures X and Y, it is evident that the model fitting exercise was a success. In both cases, the proposed model is able to generate profiles which accurately fit the responses selected by the SLINGSHOTS algorithm. Utilizing a relatively simple model with constant coefficients, it is possible to reconstruct the data with a reasonable level of accuracy. The parameters associated with each model fitting are given in **Appendix B.** While it would have been ideal if the coefficients in both fits to correspond exactly, this was not an expected result. Because of the normalization that had occurred, we could not assess whether the scale change between the two cases were consistent, and therefore the scale factors associated with the parameters cannot be directly compared. The specific features which we are looking for are the combination of the sigmoidal response, and the response that shows tolerance with respect to a chronic infusion of corticosteroids, and two similar responses under the acute that differ by their time

constants. Because, the model was able to replicate both responses, it appears that with a single model the paradoxical responses appear to have been replicated.

Discussio

The first task is to determine whether the response of the system is a consequence of the model structure, or whether the model is sufficiently flexible to fit any kind of data. Because we are running a model fitting exercise in which we try to minimize the error between the model prediction and the data, it is important to validate that the model is specific to a given response, and is not a generalized model that can fit any type of response. To run this model validation, we essentially change the input into the system. Therefore, we seek to determine whether it is possible to obtain the acute response, if we utilize a step input, or the chronic response if we utilize a bolus injection of corticosteroid.

Thus the overall operation consists of running the same fitting operation except with “incorrect” input. The result of this operation is given in **Figure**. What is evident is that these models can only replicate the observed dynamics only when the correct input is utilized. From this result it appears that the hypothetical model at least shows a degree of specificity with respect to the input. Therefore, at this point we have validated the fact that our model has some characteristic response based upon the input into the system, rather than functioning as simply a curve fitting exercise.

The second piece of validation which we need to accomplish is to determine whether the model as constructed is able to replicate the dynamics of the system in response to an input which was not used previously. Because our model was designed to handle the differences between the acute and the chronic case, by design it needs to replicate those dynamics. However, it is less clear as to whether the response of the system will be replicated if an additional input is used. Since we have already determined that the qualitative response of the model is dependent upon the architecture, the question is whether this system can replicate dynamics from a dataset which we have not previously utilized.

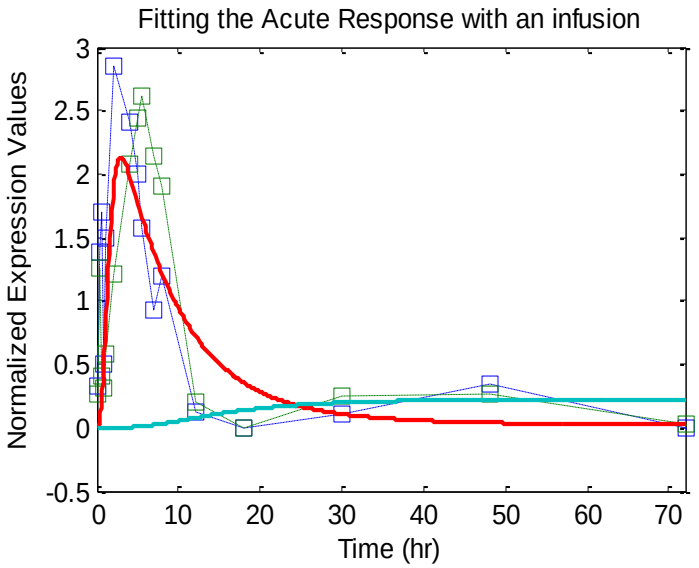


Figure 40: An attempted fit(solid) of the acute data (dotted), when then input is changed to an infusion. It appears that the dynamics of the system are determined primarily via the architecture of the model and the stimulatory input associated with an administration of corticosteroids

The specific dataset which we will use as a testing set corresponds to the response of the liver with response to the circadian rhythm of the endogenous level of corticosteroids. In the two datasets which we have based the bulk of our analysis, the adrenal glands of the animals have been removed to eliminate the effect of endogenous cortisol. This is because the adrenal glands themselves produce cortisol in a circadian manner, which would have affected the levels of corticosteroids in the system after an infusion or injection of corticosteroids. However, in the dataset which we will use for validation, the stimulus will be only that of the endogenous corticosteroid levels. Therefore, this dataset attempts to determine the effect that the circadian variations of endogenous corticosteroid levels have upon gene expression. Thus, the primary hypothesis is that genes which show circadian variations in the liver may be regulated by corticosteroids. This dataset consisted of 3 replicates over 18 different time points over 24 hours, for a total of 54 individual samples. This dataset can be found in the GEO database under accession number GDS8988. Initial analysis of this dataset suggested that the levels of corticosteroid responsive genes under the control of endogenous corticosteroids appear to follow a sinusoidal pattern[158]. Thus, the expected result of our model is that if a sinusoidal input is utilized to stimulate the liver, then the expected output will be that corticosteroid responsive genes will follow a similar pattern albeit with different magnitudes and phase shifts.

To account for this new input, we use the parameters that were obtained from the two previous fitting operations in the model, and change the input from an infusion or a bolus injection into that of a sinusoid. What we see is that if we change the levels of the endogenous level of corticosteroids to a sinusoid that the given output corresponds is also a sinusoid, which corresponds to the output which was obtained in the prior analysis of the dataset corresponding to the endogenous response to corticosteroids **Figure 41**. Thus, the response of the system to endogenous levels of corticosteroids also appears to be well handled by the proposed model.

Without having to refit the data, it appears that the architecture of the model is able to replicate the dynamics of the dataset which was not previously used. Thus, because the validation appears to be successful, this network architecture presents a reasonable hypothesis to explain the dynamics associated with transcription factor prediction.

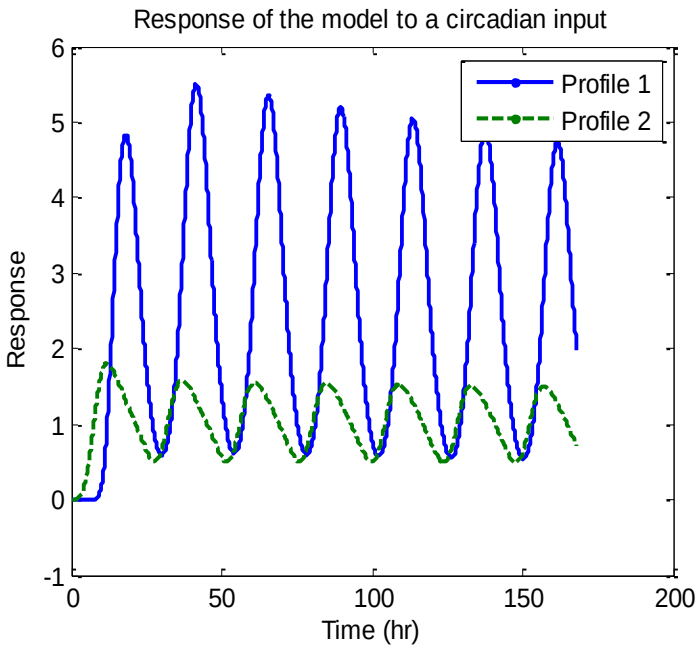


Figure 41: The response of the model to the circadian variation of endogenous corticosteroids.

It is important to note that this model was obtained without fitting the data, but only by changing the input. Furthermore, the dose was an uncalibrated sinusoid, with the initial activation of the genes set at 0. This is significant because it shows that the qualitative response of the system is dependent again only on the architecture as well as the input which we have stimulated the system with, but also that the system is stable. Making no assumptions as to the concentration of endogenous corticosteroids, nor robustly identifying the initial state, we were able to obtain a response which would stabilize over time. One of our hypotheses associated with this result pertains to the stability of our architecture. Given the fact that there exists a mass balance in the amount of receptor present in the system, we hypothesize that the system will always return to baseline once the input stimuli has been removed from the system. Thus, the occurrence of the sustained metabolic side effects may be mediated by carefully controlling a set of bolus injections of corticosteroids, such that there is sufficient time for the secondary form of the receptor to be converted back into its initial active state.

However, while the model itself appears to replicate the observed dynamics, there are a few issues which we have not accounted for, which may be important. The first issue is that of scale. At this point, we were more interested in replicating the dynamics of the signal, specifically in terms of the time constants associated with the activation/deactivation of various gene products, we have rescaled the dynamics via the z-score. However, in real biological system there are issues relating to scale, such that if the activation is under a certain threshold, downstream pathways will not be activated. In our model, we cannot directly account for this. Part of the reason for this is that we do not have sufficient experimental data to determine whether there exists an all or nothing behavior in response to corticosteroids as in the case of neuronal firing. We hypothesize that this is not the case, due to the oscillatory nature of the liver in response to the circadian rhythm of endogenous corticosteroids. However this fact cannot be

directly accounted for. Secondly, we have not taken into account issues such as time delays in the system nor the effects of stochastic processes, both of which may play an important in mediating the overall response to corticosteroids. In spite of this, we hypothesize that our initial modeling approach, presents a good initial starting point to verifying various mechanistic properties of how the corticosteroid receptor is activated, and how it in turn regulates other genes. This starting point, can then be used as the foundation for these more sophisticated models.

Concluding Remarks

The result of the model building exercise brings the Systems Biology method around in a full cycle. Starting with the initial high throughput gene expression experiments, and the incorporation of extra information in the form of the Living Cell Array, we have first managed to generate a small set of hypotheses concerning how the system should respond. Then via the creation of the model, we are essentially codifying the hypothesis in a mathematical form. This mathematical form then has certain mechanistic features which can then be tested with further experimentation. Thus, the concluding remark of this dissertation involves the codification of a hypothesis and the beginning of a proposal for future work.

The specific mechanistic consequence which we hypothesize to play an important role in the transcriptional regulation of corticosteroids involves the need for a feedback interaction involving the conversion of a monomeric form of corticosteroids into a dimeric form. These two forms then regulate two separate categories of genes, one which are involved in the regulation of inflammatory genes, and the second category which are involved in the regulation of metabolic processes. Therefore, from the results of this analysis, we can begin to propose additional experiments which can be run. The most obvious experiment which we propose is finding out whether corticosteroids themselves are active in their monomeric form. We

hypothesize that the Living Cell Array with its reporter plasmids may be well adapted to this role. Rather than utilizing the full transcription factor binding site for the glucocorticosteroid receptor as reported by TRANSFAC, we can utilize the conserved half site. If the reporter gene were constructed in this fashion, and still the anti-inflammatory response of corticosteroids are still active, then it will have provided evidence suggesting that a monomeric form of the glucocorticosteroid receptor is indeed active in the regulation of genes related to inflammation. If this mechanism could be validated, then it suggests to us the possibility of targeting the anti-inflammatory mechanisms associated with corticosteroid activity without the attendant side effects associated with prolonged corticosteroid treatment, in which the primary mode of action would be the prevention of the dimerization of the glucocorticosteroid receptor.

Thus, the application of Systems Biology to the underlying stem of study has allows us to take an experiment with a standard hypothesis, “An organism responds to external stimuli through changes in gene expression,” quantify possible dynamics, and finally formulate new hypotheses for further exploration.

Moreover, since the primary goal of Systems Biology is to create and utilize a standard set of experiments and data processing techniques, many of the methods presented in this dissertation have applications to other questions. Thus, the overall value of this dissertation is more than our elucidation of a possible mechanism behind corticosteroid activity, but also in the discrete steps that it required for us to get there. Each of these discrete steps can be applied to other problems in order to obtain initial insights for further experimentation. These problems need not be related problems such as examining the response of an organism to other drugs, but may be used to study the temporal effects of any stimulus into a biological organism provided that a suitable high throughput measurement technique can be obtained. For instance,

our proposed method for evaluating the informative nature of a dataset can be applied to any type of dataset that consists of large amount of parallel temporal data, and not just those of mRNA microarray data. The desire for generality over multiple experimental protocols was the underlying motivation for the creation of a new metric, rather than the use of the more traditional evaluation through gene ontology enrichment. We wanted a method which required only the data that was present without the incorporation of outside information.

The second algorithm presented which was a method for extracting informative motifs from the data can be applicable to any high throughput dataset in which the intrinsic relationship between different signals is important as is the underlying driving force behind the signals. Therefore, this algorithm may be of some use for the analysis of other data types such that of the stock market, in which there exists a large number of features stocks all with temporal dynamics. Likewise the result of this analysis could be used to extract underlying patterns and market movement, which may be obscured by the large number of features present.

The third algorithm can be applied to any temporal signal to evaluate how reliable a given dynamic response is. Thus, it is possible to utilize this method whenever one is interested in the temporal dynamics, but are constrained by the number of replicates within the data. While it was not done here, one possible extension would be to use the algorithm to determine the number of replicates one would need for a givens signal once the variance between replicates is known or estimated. Thus, this could be applied to other experimental system such as ELISA.

While the final algorithm is less general than the other algorithms used in this dissertation, we believe that this algorithm coupled with an expanded LCA may provide significant amounts of insight into how transcriptional signaling occurs. This method addresses many of the caveats associated with network identification, and has allowed us to establish a minimum criteria for

network identification as well as a method for assessing the response of hidden states within a system.

What these disparate steps have allowed us to do is fulfill requirements for us to examine a previously unknown system with a minimum of underlying assumption. Thus, we have been able to first assess the significance of a given dataset, isolate the portions of the data which are meaningful, and finally obtain some insight as to how these different systems are tied together. Thus by doing so, it is possible to formulate a model which serves as a hypothesis for further testing.

Appendix A

Acute Corticosteroid Ontologies

Gene	Cluster	Ontology	Ontology Category
Id1	1	Signaling	Signaling
Atp1b1	1	Response to Hypoxia	Response to Hypoxia
Fnta	1	Amino Acid Prenylation	Signaling
Ptbp1	1	mRNA Splicing	Signaling
Nolc1	1	Regulation of Protein Transport	Transport
Mpi	1	Carbohydrate Metabolism	Metabolism
Hyal2	1	Carbohydrate Metabolism	Metabolism
Slc3a2	1	Carbohydrate Metabolism	Metabolism
Slc3a2	1	Carbohydrate Metabolism	Metabolism
Aldoa	1	Glycolysis	Metabolism
Txnl1	1	Electron Transport	Metabolism
Prps1	1	Purine Base Metabolism	Metabolism
Apex1	1	DNA Repair	DNA Rpair
RGD:619726	1	Regulation of Transcription	Signaling
Hnrpab	1	Regulation of Transcription	Signaling
Hnrpk	1	RNA Processing	Signaling
Nxf1	1	mRNA Processing	Signaling
Eif2b3	1	Protein Biosynthesis	Protein Synthesis
Eif2b3	1	Protein Biosynthesis	Protein Synthesis
Tcp1	1	Protein Folding	Protein Synthesis

Hspa5	1	Protein Folding	Protein Synthesis
Hspb2	1	Protein Folding	Protein Synthesis
Ube2d2	1	Protein Modification	Protein Synthesis
Mapk6	1	Animo Acid Phosporylation	Signaling
Mapk6	1	Animo Acid Phosporylation	Signaling
Ywhah	1	Negative Regulation of Protein Kinase Activity	Signaling
Metap2	1	Proteolysis	Metabolism
Metap2	1	Proteolysis	Metabolism
RGD:621595	1	Ubiquitin Cycle	Metabolism
Tomm20	1	Protein Targeting	Signaling
Ywhag	1	Protein Targeting	Signaling
Ssr3	1	Protein Membrane Targeting	Signaling
Slc29a2	1	Transport	Transport
Slc5a3	1	Transport	Transport
Atp2a2	1	Cation Transport	Transport
Atp2a2	1	Cation Transport	Transport
Cltb	1	Neurotransmitter Transport	Transport
Npm1	1	Intracellular Protein Transport	Transport
RGD:620645	1	Apoptosis	Immune Response
Zp2	1	Cytoskeletal Organization	Cytoskeletal Organization
Dncic2	1	Microtuble based processes	Cytoskeletal Organization

Cntf	1	Neurogenesis	Neurogenesis
Pafah1b1	1	Neurogenesis	Neurogenesis
Ifrd1	1	Neurogenesis	Neurogenesis
Slc12a4	1	Regulation of Cell Cycle	Cell Cycle
Ccnd3	1	Regulation of Cell Cycle	Cell Cycle
Ccnd3	1	Regulation of Cell Cycle	Cell Cycle
Avp	1	Regulation of Blood Pressure	Regulation of Blood Pressure
Hif1a	2	Angiogenesis	Angiogenesis
Psen1	2	Angiogenesis	Angiogenesis
Prkar1a	2	Mesoderm Formation	Mesoderm Formation
Prkaa1	2	Actiavation of MAPK	Signaling
Asl	2	Urea cycle	Metabolism
Asl /// Hnrpab	2	Urea cycle	Metabolism
Hyal2	2	Carbohydrate Metabolism	Metabolism
Tat	2	gluconeogenesis	Metabolism
Mybbp1a	2	Electron Transport	Metabolism
RGD:708345	2	DNA Repair	Response to Hypoxia
Bteb1	2	Regulation of Transcription	Signaling
Bteb1	2	Regulation of Transcription	Signaling
Hsf1	2	Regulation of Transcription	Signaling
Dkc1	2	Regulation of Transcription	Signaling
SMN1	2	mRNA Processing	Signaling

Eif4e	2	Protein Biosynthesis	Metabolism
Stch	2	Protein Folding	Metabolism
Ap2b1	2	Protein Assembly	Metabolism
Ap2b1	2	Protein Assembly	Metabolism
Ywhah	2	Regulation of Protein Kinase Activity	Metabolism
Ece1	2	Proteolysis	Metabolism
Pcsk7	2	Proteolysis	Metabolism
Pcsk5	2	Proteolysis	Metabolism
Cpt1b	2	Fatty Acid Metabolism	Metabolism
Grik1	2	Transport	Transport
Atp2a2	2	Cation Transport	Transport
Arf4	2	Protein Transport	Transport
Ptma	2	Anti-apoptosis	Immune Response
Tpm4	2	Muscle Contraction	Muscle Contraction
Itgam	2	Cell Adhesion	Cell Adheion
RGD:619777	2	Cell Adhesion	Cell Adheion
Rala	2	Signal Transduction	Signaling
Slc12a4	2	Regulation of Cell Cycle	Regulation of the Cell Cycle
Bcat1	2	Metabolism	Metabolism
Kdr	3	Angiogenesis	Angiogenesis
Gchfr	3	Nitric Oxide Biosynthesis	Metabolism
Dio1	3	Metabolism	Metabolism

Dia1	3	Electron Transport	Metabolism
Dia1	3	Electron Transport	Metabolism
Dia1	3	Electron Transport	Metabolism
Cyp4f2	3	Electron Transport	Metabolism
Dia1	3	Electron Transport	Metabolism
Dia1	3	Electron Transport	Metabolism
Cyp2d9 /// Cyp2d10	3	Electron Transport	Metabolism
Maob	3	Electron Transport	Metabolism
Dao1	3	Electron Transport	Metabolism
Cat	3	Electron Transport	Metabolism
Haao	3	Regulation of Transcription	Signaling
Haao	3	Regulation of Transcription	Signaling
Thrb	3	Regulation of Transcription	Signaling
Mapk9	3	Protein Amino Acid Phosphorylation	Signaling
Kynu	3	Amino Acid Metabolism	Metabolism
Pcbd	3	L-phenylalanine metabolism	Metabolism
Gamt	3	Creatine Biosynthesis	Metabolism
Lipa	3	Lipid Metabolism	Metabolism
Gpam	3	Fatty Acid Metabolism	Metabolism
Slc40a1	3	Transport	Transport
Slc40a1	3	Transport	Transport
RGD:621430	3	Transport	Transport
Ttpa	3	Transport	Transport

Slc2a2	3	Transport	Transport
Mbl2	3	Phosphate Transport	Transport
Baat	3	Acute-Phase Response	Immune Response
Jup	3	Cell Adheion	Cell Adhesion
Ceacam1	3	Signal Transduction	Signaling
Ndrg2	3	Signal Transduction	Signaling
Adra1b	3	G-protein coupled receptor protein signaling pathway	Signaling
Rab8a	3	GTPase Signal Transduction	Signaling
Fgf1	3	Cell Cycle	Cell Cycle
Serpind1	3	Coagulation	Coagulation
Nat1	3	Metabolism	Metabolism
Enpp2	3	Nuclotide Metabolism	Metabolism
Abat	3	Gamma-aminobutyric Metabolism	Metabolism

Chronic Corticosteroid Ontologies

Gene	Cluster	Ontology	Ontology Category
Ftcd	1	amino acid metabolic process	Metabolism
Slc27a5	1	bile acid metabolic process	Metabolism
Aldob	1	fructose metabolic process	Metabolism
Slc37a4	1	glucose-6-phosphate transport	Metabolism
Gcsh	1	glycine decarboxylation via glycine cleavage system	Metabolism
Hpd	1	L-phenylalanine catabolic process	Metabolism

Gchfr	1	Metabolism	Metabolism
Fah	1	Metabolism	Metabolism
Adhfe1	1	Metabolism	Metabolism
Kif15	1	microtubule-based movement	Movement
Adrbk2	1	desensitization of G-protein coupled receptor protein	Signal Transduction
Mrpl16	1	translation	Translation
Sept2	2	Cell Cycle	Cell Cycle
Ncl	2	Cellular Growth	Cellular Growth
Hrmt1l2	2	Defense Response	Immune Response
Afp	2	Immune Response	Immune Response
Adsl_predicted	2	aerobic respiration	Metabolism
Nsun2_predicted	2	Oxidoreductase Activity	Oxidoreductase Activity
Eif2s2	2	Protein Production	Protein Production
Eif3s4	2	Protein Production	Protein Production
Eif3s9	2	translational initiation	Protein Production
Prkar1a	2	cell proliferation	Signal Transduction
Syncrip	2	RNA splicing	Signal Transduction
Slc3a2	2	amino acid transport	Transport
Sugt1	3	mitosis	Cellular Growth
Ddb1	3	DNA Repair	DNA Repair
Vezf1_predicted	3	cellular defense response	Immune Response

Ube4b_predicted	3	apoptosis	Immune Response
Bcap29	3	apoptosis	Immune Response
Zfr	3	NK T cell proliferation	Immune Response
Stch	3	Stress	Immune Response
Rabggtb	3	Protein Modification	Protein Modification
Eif5	3	regulation of translational initiation	Protein Production
Ica1	3	neurotransmitter transport	Signal Transduction
Slc33a1	3	transport	Transport
Ccpg1_predicted	4	Cell Cycle	Cell Cycle
Pold4	4	DNA replication	Cellular Growth
Brinp3	4	negative regulation of cell cycle	Cellular Growth
Sdc2	4	Cellular Signaling	Cellular Signaling
Abhd1	4	abhydrolase	Metabolism
Mccc1	4	biotin metabolic process	Metabolism
Apoa1	4	cholesterol metabolic process	Metabolism
Abo	4	carbohydrate metabolic process	Metabolism
Dpyd	4	'de novo' pyrimidine base biosynthetic process	Metabolism
Ndufa6_predicted	4	Metabolism	Metabolism
Ndrg2	4	cell differentiation	Signal Transduction
Lynx1_predicted	4	synaptic transmission, cholinergic	Signal Transduction
Hspa9a_predicted	5	anti-apoptosis	Immune Response
Psmc4	5	blastocyst development	Metabolism

Oxnad1_predicted	5	Metabolism	Metabolism
Ratsg2	5	Response to Glucose Stimulation	Metabolism
Cct3	5	chaperonin-mediated tubulin folding	Protein Production
Eif5	5	regulation of translational initiation	Protein Production
Eif3s6ip	5	Protien Production	Protein Production
Arf6	5	actin cytoskeleton organization and biogenesis	Signal Transduction
Mkks	5	sensory cilium biogenesis	Signal Transduction
Psmd4	5	fluid transport	Transport
Nfia	6	DNA replication	Cellular Growth
Serpina1	6	acute-phase response	Immune Response
Cox7b	6	electron transport	Metabolism
Ugt2b	6	metabolic process	Metabolism
Svs1	6	Metabolism	Metabolism
Sulf2	6	Metabolism	Metabolism
Ctsh	6	proteolysis	Metabolism
Hp	6	ion transport	Transport
Rbp4	6	transport	transport
Arsb	7	autophagy	Metabolism
Metap2	7	N-terminal protein amino acid modification	Protein Production
Eif4a1	7	Protein Production	Protein Production
Sfpq	7	RNA splicing	Protein Production
RGD1308469_predicted	7	translation	Protein Production

Eif2s1	7	regulation of translation	regulation of translation
Mybbp1a	7	electron transport	Signal Transduction
Abce1	7	electron transport	Transport
Masp2	8	complement activation	Immune Response
Atp5e	8	ATP biosynthetic process	Metabolism
Gpt1	8	gluconeogenesis	Metabolism
Ugt2a1	8	detection of chemical stimulus	Metabolism
Them2_predicted	8	Fatty Acid Metabolism	Metabolism
Dao1	8	metabolic process	Metabolism
Ceacam1	8	angiogenesis	Signal Transduction
Sybl1	8	intracellular protein transport	Transport
Abat	9	response to hypoxia	response to hypoxia
Cyp2d22	9	arachidonic acid metabolic process	Metabolism
As3mt	9	arsonoacetate metabolic process	Metabolism
LOC298250	9	carbohydrate metabolic process	Metabolism
Khk	9	carbohydrate metabolic process	Metabolism
Osbpl9_predicted	9	Cholesterol Metabolism	Metabolism
Cox7a2	9	electron transport	Metabolism
Ndufa2_predicted	9	generation of precursor metabolites and energy	Metabolism
Hsd17b13	9	metabolic process	Metabolism
Apoc4	10	lipid metabolic process	Metabolism

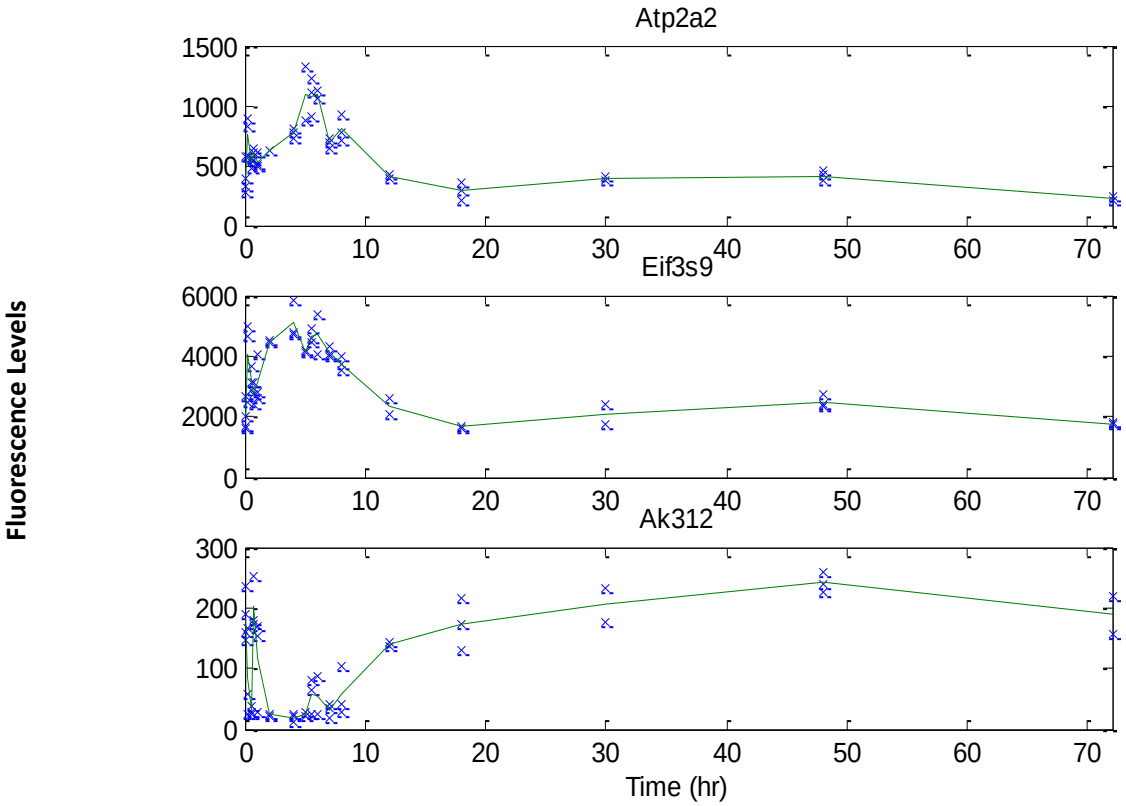
Slc2a5	10	carbohydrate metabolic process	Metabolism
Ppp1r3c	10	carbohydrate metabolic process	Metabolism
Atox1	10	Oxidative Stress	Metabolism
LOC300963	10	Protein Binding	Protein Binding
Amph1	10	Regulation of GTPase	Signal Transduction
Rpp21	10	ribonuclease P 21 subunit	Signal Transduction
Omp	10	sensory perception of smell	Signal Transduction
Gfra4	10	Receptor fortyrosine kinase signaling pathway	Signal Transduction

Appendix B

Parameters	Acute Dosing	Chronic Dosing
K_1 (D + R -> DR)	$1.94x10^{-1}$	$2.67x10^{-4}$
K_1 (DR -> D + R)	$2.08x10^{-5}$	$8.80x10^{-6}$
K_3 (DR-> DRN)	5.71	$2.4x10^{-1}$
K_4 (DRN + $mRNA_1$->R*)	3.13	1.57
K_5($mRNA_1$ Synthesis)	$2.26x10^{-1}$	$1.3x10^{-1}$
K_6 (mRNA Degradation)	$6.42x10^{-1}$	$3.4x10^{-1}$
K_7 (R*->$D_2R^*_2$)	$3.60x10^{-1}$	$3.9x10^{-2}$
K_8($D_2R^*_2$-> R*)	14.6	2.4
K_9 ($mRNA_2$Synthesis)	$1.8x10^{-2}$	2.14
K_{10} ($mRNA_2$ Degradation)	$3.84x10^{-1}$	$8x10^{-3}$
K_{11}(R*->R)	$4.02x10^{-1}$	$1.5x10^{-1}$

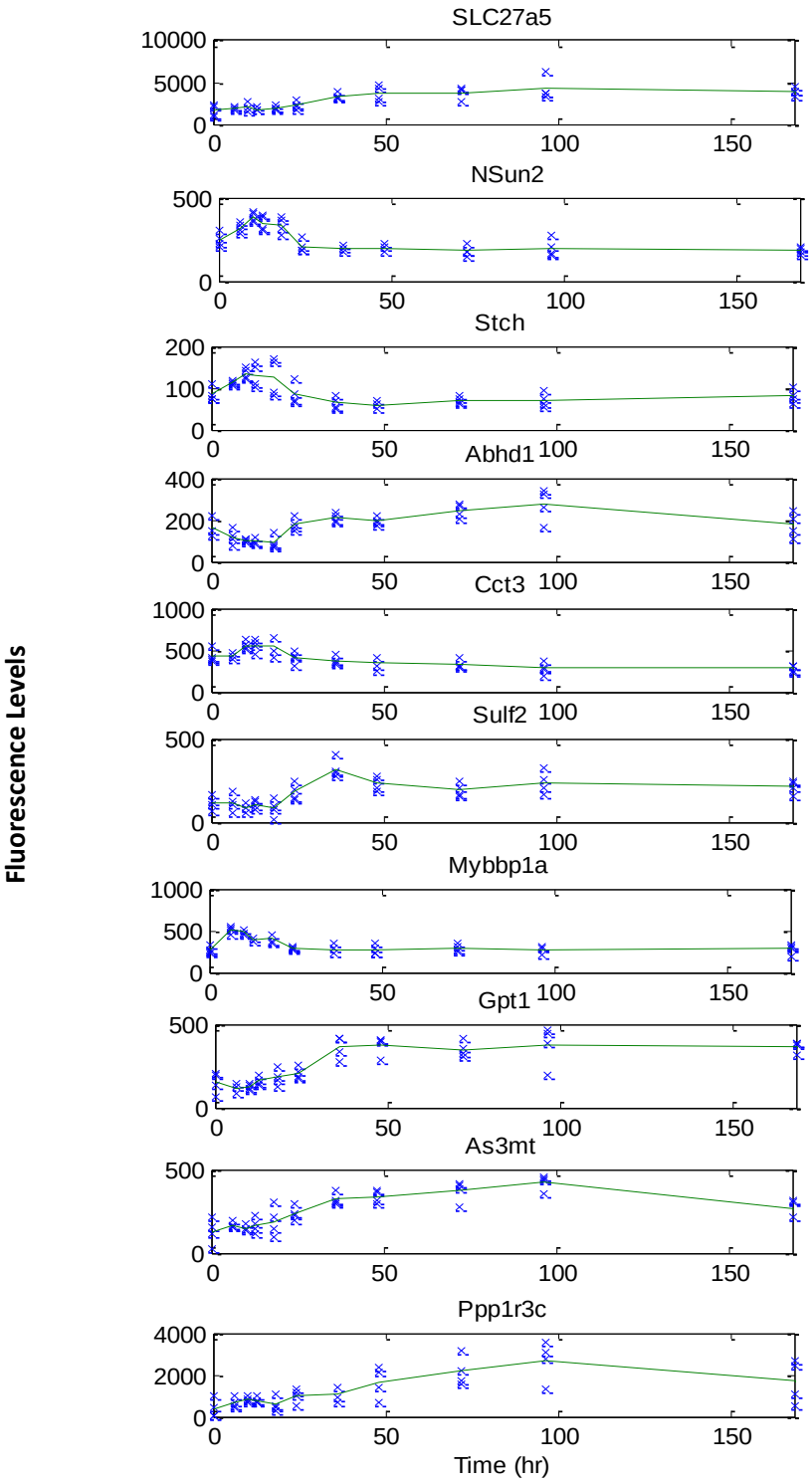
The coefficients associated with the proposed model (7.1). The coefficients are reported without units because of the qualitative nature of the model, as well as our inability to calibrate the magnitude of the responses due to ambiguities in translating the fluoresce levels between array platforms and probe sets.

Appendix C



The selected marker genes for the three clusters obtained under the acute case. Note the relatively small spread associated with the replicates

Selected Chronic Profiles



The profiles of selected marker genes under a chronic infusion of corticosteroids. Note the lack of variability within each of the replicates. Thus, it appears that these profiles have been accurately determined

Acknowledgement of Previous Publications

This dissertation incorporates significant portions of previous publications listed below:

1. Yang, E., M.L. Yarmush and I.P. Androulakis, Transcription Factor Network Reconstruction using the Living Cell Array", J. Theo. Biology Accepted (2008)
2. Yang, E., K. King, M.L. Yarmush and I.P Androulakis, Extraction of Transcriptional Signaling Networks via Globally Optimal Biclustering. Proceedings of the 5th International Conference of the Foundations of Computer-Aided Process Operations (FOCAPO), Cambridge, MA (2008)
3. Yang, E., R.R. Almon, D.C. Dubois, W.J. Jusko and I.P. Androulakis, Extracting Global System Dynamics of Corticosteroid Genomic Effects in Rat Liver. Journal of Pharmacology and Experimental Therapeutics, doi:10.1124/jpet.107.133074 (2007)
4. Foteinou, P.T., E. Yang, G.K. Saharidis, M.G. Ierapetritou and I.P. Androulakis, A systematic framework for the synthesis and analysis of regulatory networks. Journal of Global Optimization doi:10.1007/s10898-007-9266-6 (2007)
5. Yang, E., P.T. Foteinou, K.R. King, M.L. Yarmush and I.P. Androulakis, A Novel Non-overlapping Bi-clustering Algorithm for Network Generation using Living Cell Array data. Oxford Bioinformatics,23(17):2306 (2007)
6. Yang, E. and I.P. Androulakis, Assessing the Information Content of Microarray Time Series." Encyclopedia of Healthcare Information Systems IGI Global (2008)
7. Yang, E., D. Simcha, R.R. Almon, D.C. Dubois, W.J. Jusko and I.P. Androulakis, Context Specific Transcription Factor Prediction. Annals of Biomedical Engineering, 35(6):1053 (2007)
8. Yang, E., T. Maguire, M.L. Yarmush, F. Berthiaume and I.P. Androulakis, Bioinformatics Analysis of the Early Inflammatory Response in a Rat Thermal Injury Model. BMC Bioinformatics, 8:10 (2007)
9. Yang, E. and I.P. Androulakis, Information Content of Short Time Series Expression Data. Proceedings of the 28th IEEE EMBS Annual International Conference, 1:5535 (2006)

Bibliography

1. Almon, R.R., et al., *Development, Analysis, and Use of Pharmacogenomic Time Series for Pharmacokinetic/Pharmacodynamic Modeling of Multi-tissue Polygenic Responses to Corticosteroids.* Progress in Pharmacogenetics Research, 2005(in press).
2. Airla, N., et al., *Suppression of immune system genes by methylprednisolone in exacerbations of multiple sclerosis. Preliminary results.* J Neurol, 2004. **251**(10): p. 1215-9.
3. Sun, Y.N., et al., *Dose-dependence and repeated-dose studies for receptor/gene-mediated pharmacodynamics of methylprednisolone on glucocorticoid receptor down-regulation and tyrosine aminotransferase induction in rat liver.* J Pharmacokinet Biopharm, 1998. **26**(6): p. 619-48.
4. Barber, A.E., et al., *Glucocorticoid therapy alters hormonal and cytokine responses to endotoxin in man.* J Immunol, 1993. **150**(5): p. 1999-2006.
5. Jin, J.Y., et al., *Modeling of corticosteroid pharmacogenomics in rat liver using gene microarrays.* J Pharmacol Exp Ther, 2003. **307**(1): p. 93-109.
6. Samra, J.S., et al., *Effects of physiological hypercortisolemia on the regulation of lipolysis in subcutaneous adipose tissue.* J Clin Endocrinol Metab, 1998. **83**(2): p. 626-31.
7. Bae, S.C., et al., *Cost-effectiveness of low dose corticosteroids versus non-steroidal anti-inflammatory drugs and COX-2 specific inhibitors in the long-term treatment of rheumatoid arthritis.* Rheumatology (Oxford), 2003. **42**(1): p. 46-53.
8. Mager, D.E., N. Moledina, and W.J. Jusko, *Relative immunosuppressive potency of therapeutic corticosteroids measured by whole blood lymphocyte proliferation.* J Pharm Sci, 2003. **92**(7): p. 1521-5.
9. Goulding, N.J., *Corticosteroids--a case of mistaken identity?* Br J Rheumatol, 1998. **37**(5): p. 477-80.
10. Ramakrishnan, R., et al., *Fifth-generation model for corticosteroid pharmacodynamics: application to steady-state receptor down-regulation and enzyme induction patterns during seven-day continuous infusion of methylprednisolone in rats.* J Pharmacokinet Pharmacodyn, 2002. **29**(1): p. 1-24.
11. Dayneka, N.L., V. Garg, and W.J. Jusko, *Comparison of four basic models of indirect pharmacodynamic responses.* J Pharmacokinet Biopharm, 1993. **21**(4): p. 457-78.
12. Kufe, D.W., E. Frei, and J.F. Holland, *Cancer medicine 7*. 7th ed. 2006, Hamilton ; London: Elsevier. xxiii, 2328 p.
13. Almon, R.R., et al., *In vivo multi-tissue corticosteroid microarray time series available online at Public Expression Profile Resource (PEPR).* Pharmacogenomics, 2003. **4**(6): p. 791-9.
14. Almon, R.R., et al., *Gene arrays and temporal patterns of drug response: corticosteroid effects on rat liver.* Funct Integr Genomics, 2003. **3**(4): p. 171-9.
15. Mc, D.I. and M. Reich, *Corticosteroid secretion by the autotransplanted adrenal gland of the conscious sheep.* J Physiol, 1959. **147**(1): p. 33-50.
16. Oppenheim, A.V., A.S. Willsky, and S.H. Nawab, *Signals & systems*. 2nd ed. Prentice-Hall signal processing series. 1997, Upper Saddle River, N.J.: Prentice Hall. xxx, 957 p.
17. Ljung, L., *System identification : theory for the user*. 2nd ed. Prentice Hall information and system sciences series. 1999, Upper Saddle River, NJ: Prentice Hall PTR. xxii, 609 p.
18. Piroddi, L. and A. Leva, *Step response classification for model-based autotuning via polygonal curve approximation.* Journal of Process Control, 2007. **17**(8): p. 641-652.

19. Almon, R.R., D.C. Dubois, and W.J. Jusko, *A microarray analysis of the temporal response of liver to methylprednisolone: a comparative analysis of two dosing regimens.* Endocrinology, 2007. **148**(5): p. 2209-25.
20. Jenson, S.D., et al., *Validation of cDNA microarray gene expression data obtained from linearly amplified RNA.* Mol Pathol, 2003. **56**(6): p. 307-12.
21. Koscielny, S., et al., *Validation of microarray data by quantitative reverse-transcriptase polymerase chain reaction.* J Clin Oncol, 2005. **23**(36): p. 9439-40; author reply 9440.
22. Cheung, V.G., et al., *Making and reading microarrays.* Nat Genet, 1999. **21**(1 Suppl): p. 15-9.
23. Bar-Joseph, Z., *Analyzing time series gene expression data.* Bioinformatics, 2004. **20**(16): p. 2493-503.
24. Semmlow, J.L., *Biosignal and biomedical image processing : MATLAB-based applications*. 2004, New York: Marcel Dekker. xviii, 423 p.
25. Cao, D. and R. Parker, *Computational modeling of eukaryotic mRNA turnover.* Rna, 2001. **7**(9): p. 1192-212.
26. Zhu, G., et al., *Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth.* Nature, 2000. **406**(6791): p. 90-4.
27. Yang, E. and I. Androulakis. *Assessing the Information Content of Short Time Series Expression Data*. in *EMBS*. 2006. NY.
28. Fujita, A., et al., *Modeling gene expression regulatory networks with the sparse vector autoregressive model.* BMC Syst Biol, 2007. **1**: p. 39.
29. Wolfe, C.J., I.S. Kohane, and A.J. Butte, *Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks.* BMC Bioinformatics, 2005. **6**: p. 227.
30. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
31. Draghici, S., et al., *Global functional profiling of gene expression.* Genomics, 2003. **81**(2): p. 98-104.
32. Chen, G.X., et al., *Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data.* Statistica Sinica, 2002. **12**(1): p. 241-262.
33. Kannan, R., S. Vempala, and A. Vetta, *On Clusterings: Good, Bad, and Spectral.* Proceedings of the 41st Annual Symposium on the Foundation of Computer Science, 2000: p. 367-380.
34. Gibbons, F.D. and F.P. Roth, *Judging the quality of gene expression-based clustering methods using gene annotation.* Genome Res, 2002. **12**(10): p. 1574-81.
35. Karypis, G., E.-H.S. Han, and V. Kumar, *Chameleon: Hierarchical Clustering Using Dynamic Modeling.* Computers, 1999. **32**(8): p. 68--75.
36. Jayaraman, A., M.L. Yarmush, and C.M. Roth, *Evaluation of an in vitro model of hepatic inflammatory response by gene expression profiling.* Tissue Eng, 2005. **11**(1-2): p. 50-63.
37. Baldi, P., *DNA Microarrays and Gene Expression : From Experiments to Data Analysis and Modeling*. 2002, Cambridge: CAMBRIDGE UNIV PRESS.
38. Datta, S. and S. Datta, *Comparisons and validation of statistical clustering techniques for microarray gene expression data.* Bioinformatics, 2003. **19**(4): p. 459-66.
39. Choi, J.K., et al., *Differential coexpression analysis using microarray data and its application to human cancer.* Bioinformatics, 2005. **21**(24): p. 4348-55.
40. Churchill, G.A., *Using ANOVA to analyze microarray data.* Biotechniques, 2004. **37**(2): p. 173-5, 177.

41. Fox, R.J. and M.W. Dimmic, *A two-sample Bayesian t-test for microarray data.* BMC Bioinformatics, 2006. **7**: p. 126.
42. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
43. Nielsen, H.B., L. Gautier, and S. Knudsen, *Implementation of a gene expression index calculation method based on the PDNN model.* Bioinformatics, 2005. **21**(5): p. 687-8.
44. Dupuy, A. and R.M. Simon, *Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting.* J Natl Cancer Inst, 2007. **99**(2): p. 147-57.
45. Dudoit, S., et al., *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.* Statistica Sinica, 2002. **12**(1): p. 111-139.
46. Izumo, M., et al., *Quantitative analyses of circadian gene expression in mammalian cell cultures.* PLoS Comput Biol, 2006. **2**(10): p. e136.
47. Klebanov, L., et al., *Statistical methods and microarray data.* Nat Biotechnol, 2007. **25**(1): p. 25-6; author reply 26-7.
48. Keogh, E., J. Lin, and A. Fu. *HOT SAX: Efficiently finding the most unusual time series subsequences*. in *5th IEEE International Conference on Data Mining*. 2005.
49. Ernst, J. and Z. Bar-Joseph, *STEM: a tool for the analysis of short time series gene expression data.* BMC Bioinformatics, 2006. **7**: p. 191.
50. Yang, E., et al., *An Integrative Systems Biology Approach for Analyzing Liver Hypermetabolism.* Proceedings of the Joint 9th International Symposium, Processing Systems Engineering and 16th European Symposium, 2006.
51. Spiegel, M.R. and L.J. Stephens, *Schaum's outline of theory and problems of statistics*. 3rd ed. 1999, New York: McGraw-Hill. xvii, 538 p.
52. Rassokhin, D.N. and D.K. Agrafiotis, *Kolmogorov-Smirnov statistic and its application in library design.* J Mol Graph Model, 2000. **18**(4-5): p. 368-82.
53. NIST. *e-Handbook of Statistical Methods*. 1998 [cited; Available from: http://www.itl.nist.gov/div898/handbook/.
54. Ramakrishnan, R., et al., *Pharmacodynamics and pharmacogenomics of methylprednisolone during 7-day infusions in rats.* J Pharmacol Exp Ther, 2002. **300**(1): p. 245-56.
55. Meduri, G.U., et al., *Methylprednisolone infusion in early severe ARDS: results of a randomized controlled trial.* Chest, 2007. **131**(4): p. 954-63.
56. Perlstein, I., et al., *A signal transduction pharmacodynamic model of the kinetics of the parasympathomimetic activity of low-dose scopolamine and atropine in rats.* J Pharm Sci, 2002. **91**(12): p. 2500-10.
57. de Blaauw, I., et al., *De novo glutamine synthesis induced by corticosteroids in vivo in rats is secondary to weight loss.* Clin Nutr, 2004. **23**(5): p. 1035-42.
58. Tu, Y., G. Stolovitzky, and U. Klein, *Quantitative noise analysis for gene expression microarray experiments.* Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14031-6.
59. Mutch, D.M., et al., *The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data.* BMC Bioinformatics, 2002. **3**: p. 17.
60. Kearns, M. and D. Ron, *Algorithmic stability and sanity-check bounds for leave-one-out cross-validation.* Neural Computation, 1999. **11**(6): p. 1427-1453.
61. Luan, Y. and H. Li, *Clustering of time-course gene expression data using a mixed-effects model with B-splines.* Bioinformatics, 2003. **19**(4): p. 474-82.
62. Ramoni, M.F., P. Sebastiani, and I.S. Kohane, *Cluster analysis of gene expression dynamics.* Proc Natl Acad Sci U S A, 2002. **99**(14): p. 9121-6.

63. Li, S., M.J. Becich, and J. Gilbertson, *Microarray data mining using gene ontology.* Medinfo, 2004. **11**(Pt 2): p. 778-82.
64. Pavlidis, P., *Using ANOVA for gene selection from microarray studies of the nervous system.* Methods, 2003. **31**(4): p. 282-9.
65. Affymetrix (2003) *Array Design and Performance of the GeneChip® Rat Expression Set 230*. **Volume**,
66. Moller-Levet, C.S., et al., *Clustering of unevenly sampled gene expression time-series data.* Fuzzy Sets and Systems, 2005. **152**(1): p. 49-66.
67. Ren, B., et al., *Genome-wide location and function of DNA binding proteins.* Science, 2000. **290**(5500): p. 2306-9.
68. Wasserman, W.W. and A. Sandelin, *Applied bioinformatics for the identification of regulatory elements.* Nat Rev Genet, 2004. **5**(4): p. 276-87.
69. Matys, V., et al., *TRANSFAC: transcriptional regulation, from patterns to profiles.* Nucleic Acids Res, 2003. **31**(1): p. 374-8.
70. Fang, F. and M. Blanchette, *FootPrinter3: phylogenetic footprinting in partially alignable sequences.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W617-20.
71. Neph, S. and M. Tompa, *MicroFootPrinter: a tool for phylogenetic footprinting in prokaryotic genomes.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W366-8.
72. Allende, M.L., et al., *Cracking the genome's second code: Enhancer detection by combined phylogenetic footprinting and transgenic fish and frog embryos.* Methods, 2006. **39**(3): p. 212-9.
73. Corcoran, D.L., E. Feingold, and P.V. Benos, *FOOTER: a web tool for finding mammalian DNA regulatory regions using phylogenetic footprinting.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W442-6.
74. von Bubnoff, A., et al., *Phylogenetic footprinting and genome scanning identify vertebrate BMP response elements and new target genes.* Dev Biol, 2005. **281**(2): p. 210-26.
75. Bigelow, H.R., et al., *CisOrtho: a program pipeline for genome-wide identification of transcription factor target genes using phylogenetic footprinting.* BMC Bioinformatics, 2004. **5**: p. 27.
76. Berezikov, E., et al., *CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting.* Genome Res, 2004. **14**(1): p. 170-8.
77. Blanchette, M. and M. Tompa, *FootPrinter: A program designed for phylogenetic footprinting.* Nucleic Acids Res, 2003. **31**(13): p. 3840-2.
78. Jiao, K., et al., *Phylogenetic footprinting reveals multiple regulatory elements involved in control of the meiotic recombination gene, REC102.* Yeast, 2002. **19**(2): p. 99-114.
79. Almon, R.R., et al., *Corticosteroid-regulated genes in rat kidney: mining time series array data.* Am J Physiol Endocrinol Metab, 2005. **289**(5): p. E870-82.
80. Sandelin, A., W.W. Wasserman, and B. Lenhard, *ConSite: web-based prediction of regulatory elements using cross-species comparison.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W249-52.
81. Allocco, D.J., I.S. Kohane, and A.J. Butte, *Quantifying the relationship between co-expression, co-regulation and gene function.* BMC Bioinformatics, 2004. **5**: p. 18.
82. Dieterich, C., et al., *CORG: a database for COmparative Regulatory Genomics.* Nucleic Acids Res, 2003. **31**(1): p. 55-7.
83. Zhao, F., et al., *TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies.* Nucleic Acids Res, 2005. **33**(Database issue): p. D103-7.

84. Xuan, Z., et al., *Genome-wide promoter extraction and analysis in human, mouse, and rat.* Genome Biol, 2005. **6**(8): p. R72.
85. Dieterich, C., et al., *Annotating regulatory DNA based on man-mouse genomic comparison.* Bioinformatics, 2002. **18 Suppl 2**: p. S84-90.
86. Madan Babu, M., S.A. Teichmann, and L. Aravind, *Evolutionary dynamics of prokaryotic transcriptional regulatory networks.* J Mol Biol, 2006. **358**(2): p. 614-33.
87. Rodriguez-Caso, C., M.A. Medina, and R.V. Sole, *Topology, tinkering and evolution of the human transcription factor network.* Febs J, 2005. **272**(24): p. 6423-34.
88. Jeong, H., et al., *The large-scale organization of metabolic networks.* Nature, 2000. **407**(6804): p. 651-4.
89. Balaji, S., et al., *Uncovering a hidden distributed architecture behind scale-free transcriptional regulatory networks.* J Mol Biol, 2006. **360**(1): p. 204-12.
90. Alter, O. and G.H. Golub, *Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription.* Proc Natl Acad Sci U S A, 2004. **101**(47): p. 16577-82.
91. Kato, M., et al., *Identifying combinatorial regulation of transcription factors and binding motifs.* Genome Biol, 2004. **5**(8): p. R56.
92. Gao, F., B.C. Foat, and H.J. Bussemaker, *Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data.* BMC Bioinformatics, 2004. **5**: p. 31.
93. Boulesteix, A.L. and K. Strimmer, *Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach.* Theor Biol Med Model, 2005. **2**: p. 23.
94. Yeung, M.K., J. Tegner, and J.J. Collins, *Reverse engineering gene networks using singular value decomposition and robust regression.* Proc Natl Acad Sci U S A, 2002. **99**(9): p. 6163-8.
95. Bussemaker, H.J., H. Li, and E.D. Siggia, *Regulatory element detection using correlation with expression.* Nat Genet, 2001. **27**(2): p. 167-71.
96. Liao, J.C., et al., *Network component analysis: reconstruction of regulatory signals in biological systems.* Proc Natl Acad Sci U S A, 2003. **100**(26): p. 15522-7.
97. Tran, L.M., et al., *gNCA: A framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation.* Metabolic Engineering, 2005. **7**(2): p. 128-141.
98. Kao, K.C., et al., *Network component analysis of Escherichia coli transcriptional regulation.* Abstracts of Papers of the American Chemical Society, 2004. **227**: p. U216-U217.
99. Kao, K.C., et al., *Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis.* Proc Natl Acad Sci U S A, 2004. **101**(2): p. 641-6.
100. Kao, K.C., L.M. Tran, and J.C. Liao, *A global regulatory role of gluconeogenic genes in Escherichia coli revealed by transcriptome network analysis.* J Biol Chem, 2005. **280**(43): p. 36079-87.
101. Sun, N., R.J. Carroll, and H. Zhao, *Bayesian error analysis model for reconstructing transcriptional regulatory networks.* Proc Natl Acad Sci U S A, 2006. **103**(21): p. 7988-93.
102. Salgado, H., et al., *RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12.* Nucleic Acids Res, 2001. **29**(1): p. 72-4.
103. Brooke, A., D. Kendrick, Meeraus, *GAMS: A Users Guide.* Scientific Press, Palo Alto, Calif., 1988.

104. Harbison, C.T., et al., *Transcriptional regulatory code of a eukaryotic genome.* Nature, 2004. **431**(7004): p. 99-104.
105. Lee, T.I., et al., *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science, 2002. **298**(5594): p. 799-804.
106. Stormo, G.D. and D.S. Fields, *Specificity, free energy and information content in protein-DNA interactions.* Trends Biochem Sci, 1998. **23**(3): p. 109-13.
107. Drazinic, C.M., et al., *Activation mechanism of the multifunctional transcription factor repressor-activator protein 1 (Rap1p).* Mol Cell Biol, 1996. **16**(6): p. 3187-96.
108. Boscolo, R., et al., *A generalized framework for network component analysis.* IEEE/ACM Trans Comput Biol Bioinform, 2005. **2**(4): p. 289-301.
109. Biegler, L.T., I.E. Grossmann, and A.W. Westerberg, *Systematic Methods of Chemical Process Design*. 1997: Prentice Hall.
110. Oh, M.K., et al., *Global expression profiling of acetate-grown Escherichia coli.* J Biol Chem, 2002. **277**(15): p. 13175-83.
111. Pournara, I. and L. Wernisch, *Factor analysis for gene regulatory networks and transcription factor activity profiles.* BMC Bioinformatics, 2007. **8**: p. 61.
112. Covert, M.W., et al., *Integrating high-throughput and computational data elucidates bacterial networks.* Nature, 2004. **429**(6987): p. 92-6.
113. Winzeler, E.A., et al., *Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis.* Science, 1999. **285**(5429): p. 901-6.
114. Segal, E., et al., *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.* Nat Genet, 2003. **34**(2): p. 166-76.
115. Sun, W., T. Yu, and K.C. Li, *Detection of eQTL modules mediated by activity levels of transcription factors.* Bioinformatics, 2007. **23**(17): p. 2290-7.
116. Sasse, J., et al., *Mutational analysis of acute-phase response factor/Stat3 activation and dimerization.* Mol Cell Biol, 1997. **17**(8): p. 4677-86.
117. Renard, P., et al., *Development of a sensitive multi-well colorimetric assay for active NFkappaB.* Nucleic Acids Res, 2001. **29**(4): p. E21.
118. Gardner, T.S., et al., *Inferring genetic networks and identifying compound mode of action via expression profiling.* Science, 2003. **301**(5629): p. 102-5.
119. King, K.R., et al., *A high-throughput microfluidic real-time gene expression living cell array.* Lab Chip, 2007. **7**(1): p. 77-85.
120. King, K.R., et al., *Microfluidic flow-encoded switching for parallel control of dynamic cellular microenvironments.* Lab Chip, 2008. **8**(1): p. 107-16.
121. Somorjai, R.L., B. Dolenko, and R. Baumgartner, *Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions.* Bioinformatics, 2003. **19**(12): p. 1484-91.
122. Yang, E., et al., *A novel non-overlapping bi-clustering algorithm for network generation using living cell array data.* Bioinformatics, 2007. **23**(17): p. 2306-13.
123. Hemberg, M. and M. Barahona, *Perfect sampling of the master equation for gene regulatory networks.* Biophys J, 2007. **93**(2): p. 401-10.
124. Wieder, K.J., et al., *Optimization of reporter cells for expression profiling in a microfluidic device.* Biomed Microdevices, 2005. **7**(3): p. 213-22.
125. Thompson, D.M., et al., *Dynamic gene expression profiling using a microfabricated living cell array.* Anal Chem, 2004. **76**(14): p. 4098-103.
126. Thompson, D.A., et al., *Dynamic gene expression profiling using a microfabricated living cell array.* Annal. Chem., 2004. **76**: p. 4098-4103.

127. Cheng, Y. and G.M. Church, *Biclustering of expression data.* Proc Int Conf Intell Syst Mol Biol, 2000. **8**: p. 93-103.
128. Lin, J., et al. *A symbolic Representation of Time series, with Implication for Streaming Algorithms*. in *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery`*. 2003. San Diego, CA: ACM.
129. Rahnenfuhrer, J., et al., *Calculating the statistical significance of changes in pathway activity from gene expression data.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article16.
130. Zhang, D.J.a.A., *Cluster Analysis for Gene Expression Data: A Survey.* Jiang, D. X., and Zhang, A. Cluster Analysis for Gene Expression Data: A Survey. Technical Report 2002-06, State University of New York at Buffalo, 2002., 2002.
131. Liu, X. and L. Wang, *Computing the maximum similarity bi-clusters of gene expression data.* Bioinformatics, 2007. **23**(1): p. 50-6.
132. Mjolsness, E., D.H. Sharp, and J. Reinitz, *A connectionist model of development.* J Theor Biol, 1991. **152**(4): p. 429-53.
133. D'Haeseleer, P., et al., *Linear modeling of mRNA expression levels during CNS development and injury.* Pac Symp Biocomput, 1999: p. 41-52.
134. Guthke, R., et al., *Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection.* Bioinformatics, 2005. **21**(8): p. 1626-34.
135. Dasika, M.S., A. Gupta, and C.D. Maranas, *A mixed integer linear programming (MILP) framework for inferring time delay in gene regulatory networks.* Pac Symp Biocomput, 2004: p. 474-85.
136. Schmitt, W.A., Jr., R.M. Raab, and G. Stephanopoulos, *Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data.* Genome Res, 2004. **14**(8): p. 1654-63.
137. Rice, J. and M. Rosenblatt, *Smoothing Splines: Regression, Derivatives and Deconvolution.* The Annals of Statistics, 1983. **11**(1): p. 141-156.
138. Haverty, P.M., U. Hansen, and Z. Weng, *Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification.* Nucleic Acids Res, 2004. **32**(1): p. 179-88.
139. Kauffman, S., et al., *Random Boolean network models and the yeast transcriptional network.* Proc Natl Acad Sci U S A, 2003. **100**(25): p. 14796-9.
140. Leva, A. and L. Piroddi, *Model-specific autotuning of classical regulators: A neural approach to structural identification.* Control Engineering Practice, 1996. **4**(10): p. 1381-1391.
141. Rao, C.V. and A.P. Arkin, *Control motifs for intracellular regulatory networks.* Annu Rev Biomed Eng, 2001. **3**: p. 391-419.
142. Mangan, S. and U. Alon, *Structure and function of the feed-forward loop network motif.* Proc Natl Acad Sci U S A, 2003. **100**(21): p. 11980-5.
143. Milo, R., et al., *Network motifs: simple building blocks of complex networks.* Science, 2002. **298**(5594): p. 824-7.
144. Zhu, X., M. Gerstein, and M. Snyder, *Getting connected: analysis and principles of biological networks.* Genes Dev, 2007. **21**(9): p. 1010-24.
145. Dudbridge, F. and B.P. Koeleman, *Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies.* Am J Hum Genet, 2004. **75**(3): p. 424-35.
146. Saile, B., et al., *Interferon-gamma acts proapoptotic on hepatic stellate cells (HSC) and abrogates the antiapoptotic effect of interferon-alpha by an HSP70-dependant pathway.* Eur J Cell Biol, 2004. **83**(9): p. 469-76.

147. Campbell, K.J. and N.D. Perkins, *Post-translational modification of RelA(p65) NF-kappaB.* Biochem Soc Trans, 2004. **32**(Pt 6): p. 1087-9.
148. Sass, G., K. Koerber, and G. Tiegs, *TNF tolerance and cytotoxicity in the liver: the role of interleukin-1beta, inducible nitric oxide-synthase and heme oxygenase-1 in D-galactosamine-sensitized mice.* Inflamm Res, 2002. **51**(5): p. 229-35.
149. Xie, Y., et al., *Heat shock factor 1 represses transcription of the IL-1beta gene through physical interaction with the nuclear factor of interleukin 6.* J Biol Chem, 2002. **277**(14): p. 11802-10.
150. Saklatvala, J., P. Kaur, and F. Guesdon, *Phosphorylation of the small heat-shock protein is regulated by interleukin 1, tumour necrosis factor, growth factors, bradykinin and ATP.* Biochem J, 1991. **277 (Pt 3)**: p. 635-42.
151. Yamada, Y., et al., *Initiation of liver growth by tumor necrosis factor: deficient liver regeneration in mice lacking type I tumor necrosis factor receptor.* Proc Natl Acad Sci U S A, 1997. **94**(4): p. 1441-6.
152. Hatzigeorgiou, D.E., et al., *IL-6 down-modulates the cytokine-enhanced antileishmanial activity in human macrophages.* J Immunol, 1993. **151**(7): p. 3682-92.
153. Levy, D.E., *Interferon induction of gene expression through the Jak-Stat pathway.* Seminars in Virology, 1995. **6**(3): p. 181-189.
154. Hazra, A., et al., *Assessing the dynamics of nuclear glucocorticoid-receptor complex: adding flexibility to gene expression modeling.* J Pharmacokinet Pharmacodyn, 2007. **34**(3): p. 333-54.
155. Reichardt, H.M., et al., *DNA binding of the glucocorticoid receptor is not essential for survival.* Cell, 1998. **93**(4): p. 531-41.
156. Segard-Maurel, I., et al., *Glucocorticosteroid Receptor Dimerization Investigated by Analysis of Receptor Binding to Glucocorticosteroid Responsive Elements Using a Monomer-Dimer Equilibrium Model.* Biochemistry, 1996. **35**(5): p. 1634-1642.
157. La Baer, J. and K.R. Yamamoto, *Analysis of the DNA-binding affinity, sequence specificity and context dependence of the glucocorticoid receptor zinc finger region.* J Mol Biol, 1994. **239**(5): p. 664-88.
158. Yang, E., et al., *Extracting global system dynamics of corticosteroid genomic effects in rat liver.* J Pharmacol Exp Ther, 2008. **324**(3): p. 1243-54.

Curriculum Vita

Education

- Rutgers, The State University of New Jersey **Piscataway, NJ**

  Ph.D., Biomedical Engineering (GPA: 3.855/4.000) **2004 – present**

  – Dissertation: A Systems Biology Approach for Elucidating the Underlying Mechanism behind Corticosteroid Activity
  – Teaching Assistant, 2005-2006
  – Relevant Coursework:

Computer Appl in Biomedical Engineering	Nonlinear Programming
Mol & Cell Bioengineering	Biomechanics & Biomaterials
Biocontrols & Computer Modeling	Biosignal Processing

- Johns Hopkins University **Baltimore, MD**

  BS in Biomedical Engineering/Computer Science (GPA 3.3/4.0) **1999-2003**

– Relevant Coursework:

Biochemistry	Physiological Foundations
Computer Integrated Surgery	Models of the Cardiac Myocyte
Database Systems	

Research Experience: Non-dissertation related

- Rutgers University: Department of Biomedical Engineering Piscataway, NJ
  Graduate Assistant Researcher 2004 - Present
  – miNCA – mixed integer Network Component Analysis: Creation of Hypothetical Gene Regulatory Networks
  – Diffusion Tensor Imaging – Globally Optimal Fiber Tracking
- Johns Hopkins University: Dept. of Molecular Biology and Genetics Baltimore, MD
  Undergraduate Research Assistant 1999-2003
  – Gene Analysis Package Written in Java
  – Genomic Data Analysis Software for Yeast utilizing Zope and Postgres
  – Microarray Analysis Package for Linux

Awards/Fellowships

1. IGERT (Integrative Graduate Education and Research Traineeship) Fellowship/Scholarship (2007)
2. Dorothy Leeds Fellowship/Scholarship (2006)
3. GAANN (Graduate Assistance in Areas of National Need) Fellowship (2004)

Technical Skills

- Programming Languages: C++, Java/Swing, Python, HTML, VBA, Visual Basic, Perl, JSP, CGI, ASP, SQL, PHP, GAMS, XML
- Distributed Systems: High Performance Distributed Programming with MPI
- Systems Programming: Unix Application/System/Network Programming
- Operating Systems [Programming and Administering]: OS X, Linux, AIX, Win32 environments, Solaris
- Misc: LaTeX, Eliza, Cell Culture, Microscopy

Peer Reviewed Journal Publications and Conference Proceedings

1. Yang, E., M.L. Yarmush and I.P. Androulakis, Transcription Factor Network Reconstruction using the Living Cell Array", J. Theo. Biology Accepted (2008)
2. Tung, N.T., E. Yang, and I.P. Androulakis, Machine Learning in Gene Promoter Identification", Machine Learning Research Program, NOVA Science Publishers (2008)
3. Foteinou, P.T., E. Yang, and I.P. Androulakis, Networks, Biology and Systems Engineering: A Case study in Inflammation", Proceedings of the 5th International Conference on Foundations of Computer Aided Process Operations (FOCAPO), Cambridge, MA (July 2008)
4. Yang, E., K. King, M.L. Yarmush and I.P Androulakis, Extraction of Transcriptional Signaling Networks via Globally Optimal Biclustering. Proceedings of the 5th International Conference of the Foundations of Computer-Aided Process Operations (FOCAPO), Cambridge, MA (2008)
5. Yang, E., T.J. Maguire, M.L. Yarmush, F. Berthiaume and I.P. Androulakis, Identification of Regulatory Mechanisms of the Hepatic Response to Thermal Injury. Comp. Chem. Eng., 32(1):356 (2008)
6. Yang, E., R.R. Almon, D.C. Dubois, W.J. Jusko and I.P. Androulakis, Extracting Global System Dynamics of Corticosteroid Genomic Effects in Rat Liver. Journal of Pharmacology and Experimental Therapeutics, doi:10.1124/jpet.107.133074 (2007)
7. Foteinou, P.T., E. Yang, G.K. Saharidis, M.G. Ierapetritou and I.P. Androulakis, A systematic framework for the synthesis and analysis of regulatory networks. Journal of Global Optimization doi:10.1007/s10898-007-9266-6 (2007)
8. Yang, E., P.T. Foteinou, K.R. King, M.L. Yarmush and I.P. Androulakis, A Novel Non-overlapping Bi-clustering Algorithm for Network Generation using Living Cell Array data. Oxford Bioinformatics,23(17):2306 (2007)
9. Yang, E. and I.P. Androulakis, Assessing the Information Content of Microarray Time Series." Encyclopedia of Healthcare Information Systems IGI Global (2008)
10. Androulakis, I.P. and E. Yang, \Selection of maximally informative genes", (Accepted) Encyclopedia of Optimization, 2nd Edition (C.A. Floudas and P. Pardalos, Editors), Springer Editions (2007)
11. Yang, E., T. Maguire, M.L. Yarmush and I.P. Androulakis, Informative Gene Selection and Design of Regulatory Networks Using Integer Optimization. Comp. Chem. Eng., doi:10.1016/j.compchemeng.2007.01.009 (2007)
12. Androulakis, I.P., E. Yang, R.R. Almon, D.C. Dubois and W.J. Jusko, Analysis of Time-Series Gene Expression Data: Methods, Challenges and Opportunities. Annual Review Biomedical Engineering, 9:205 (2007)
13. Yang, E., D. Simcha, R.R. Almon, D.C. Dubois, W.J. Jusko and I.P. Androulakis, Context Specific Transcription Factor Prediction. Annals of Biomedical Engineering, 35(6):1053 (2007)
14. Yang, E., T. Maguire, M.L. Yarmush, F. Berthiaume and I.P. Androulakis, Bioinformatics Analysis of the Early Inflammatory Response in a Rat Thermal Injury Model. BMC Bioinformatics, 8:10 (2007)
15. Yang, E. and I.P. Androulakis, Information Content of Short Time Series Expression Data. Proceedings of the 28th IEEE EMBS Annual International Conference, 1:5535 (2006)
16. Yang, E., F. Berthiaume, M. Yarmush and I.P. Androulakis, An Integrative Systems Biology Approach for Analyzing Liver Hypermetabolism. Proceedings of the Joint 9th Int. Symp. Process Systems Engineering and 16th European Symp. Computer Aided Process Engineering, Garmisch-Partenkirchen, Germany (2006)

Conference Presentations and Posters

1. Yang, E., P.T. Foteinou, K.R. King, M.L. Yarmush, I.P. Androulakis, Extraction of Transcriptional Signaling Networks via Globally Optimal Biclustering. Aiche National Meeting Salt Lake City, November 2007
2. Yang, E., I.P. Androulakis, A Non-independent Model for Transcription Factor Binding. Aiche National Meeting Salt Lake City, November 2007
3. Yang, E., P.T. Foteinou, K.R. King, M.L. Yarmush, I.P. Androulakis, Extraction of Transcriptional Signaling Networks via Globally Optimal Biclustering. BMES Los Angeles, September 2007
4. Yang, E., I.P. Androulakis, A Non-independent Model for Transcription Factor Binding. BMES Fall Meeting Los Angeles, September 2007
5. Ho, Eric E. Yang, S. Gunderson, I.P. Androulakis, Degenerative Sequence Motifs Identification. AIChE National Meeting, San Francisco, November 2006
6. Yang, E., T.J. Maguire, F. Berthiaume, M.L. Yarmush, I.P. Androulakis, Bioinformatic Profiling of Short Term Liver Response to Thermal Injury. AIChE National Meeting, San Francisco, November 2006
7. Yang, E., R.R. Almon, D. Dubois, W. Jusko, I.P. Androulakis, Analysis of Corticosteroid Effects on Rat Liver. BMES Fall Meeting, Chicago, November 2006
8. Yang, E., D. Simcha, R.R. Almon, D. Dubois, W. Jusko, I.P. Androulakis, Context Specific Transcription Factor Prediction via SLINGSHOTS. BMES Fall Meeting, Chicago, November 2006

9. Yang, E., and I.P. Androulakis, Information Content of Short Time Series Expression Data IEEE EMBS Annual International Conference, New York 2006
10. Yang, E., I.P. Androulakis, Selection of Informative Genes in Time-Course Gene Expression Data. AIChE National Meeting, San Francisco, November 2006
11. Yang, E., F. Berthiaume, M. Yarmush, I.P. Androulakis, An integrative systems biology approach for analyzing liver hypermetabolism. Joint 9th Int. Symp. Process Systems Engineering and 16th European Symp. Computer Aided Process Engineering, Garmisch-Partenkirchen, Germany, July 2006
12. Yang, E., C.M. Roth, I.P. Androulakis, Mixed Integer Reformulations of Network Component Analysis. AIChE National Meeting, Cincinnati, November 2005
13. Yang, E., C.M. Roth, I.P. Androulakis, A new approach for the analysis of temporal gene expression data. AIChE National Meeting, Cincinnati, November 2005
14. Yang, E., C.M. Roth, I.P. Androulakis, Selecting maximally informative genes to enable temporal expression profiling analysis. Proceedings of Foundations of Systems Biology in Engineering, Santa Barbara, CA, August 2005