JPET Fast Forward. Published on December 17, 2007 as DOI: 10.1124/jpet.107.133074

JPET #133074

Extracting Global System Dynamics of Corticosteroid Genomic Effects in Rat Liver

E. Yang, R.R. Almon, D.C. DuBois, W.J. Jusko and I.P. Androulakis

Biomedical Engineering Department, Rutgers University, Piscataway, NJ 08854 (E.Y., IPA),

Department of Biological Sciences, SUNY Buffalo, Buffalo, NY 14260 (R.R.A, D.C.D)

Department of Pharmaceutical Sciences, SUNY Buffalo, NY 14260 (R.R.A., D.C.D., W.J.J)

Chemical & Biochemical Engineering Department, Rutgers University, Piscataway, NJ 08854 (I.P.A.)

New York State Center of Excellence in Bioinformatics and Life Sciences (R.R.A., W.J.J)

JPET #133074

Running title: Extraction of global gene dynamics

Corresponding Author:

Ioannis P. Androulakis

Biomedical Engineering Department

Rutgers University

599 Taylor Road, Piscataway, NJ 08544

Tel: 732 445 4500 x6212

Fax: 732 445 3753

e-mail: yannis@rci.rutgers.edu

Text Pages: 28 pages

Tables: 0

Figures: 8

References: 23

Abstract: 163 words

Introduction: 743 words

Discussion: 1436 words

Section: Hepatic

Abbreviations:

GR, glucocorticosteroid receptor,; MPL, Methylprednisolone; SNR, Signal to Noise Ratio; ADX, adrenalectomized;

Abstract

One of the challenges in constructing biologic models involves resolving meaningful data patterns from which the mathematical models will be generated. For models that describe the change of mRNA in response to drug administration, questions exist whether the correct genes have been selected given the myriad transcriptional effects that may occur. Oftentimes different algorithms will select or cluster different groups of genes from the same dataset. A new approach was developed that focuses on identifying the underlying global dynamics of the system instead of selecting individual genes. The procedure was applied to microarray genomic data obtained from rat liver following a large single dose of methylprednisolone in 52 adrenalectomized rats. Twelve clusters of at least 30 genes each were selected reflecting the major changes over time. This method along with isolating the underlying dynamics of the system also extracts and clusters the genes which make up this global dynamic for further analysis as to the contributions of specific mechanisms affected by the drug.

Introduction

Corticosteroids are synthetic glucocorticoids used therapeutically for their potent anti-inflammatory, anti-proliferative and immunosuppressive effects (Cronstein et al., 1992; Chikanza, 2002). They have a low therapeutic index because of the wide ranging adverse consequences of their prolonged use, including hyperglycemia, dyslipidemia, muscle wasting, hypertension, nephropathy, fatty liver, and an increased risk of atherosclerosis (Jusko WJ, 2005). They cause the liver to synthesize and release glucose within the context of steroid-induced insulin resistance thus resulting in chronic hyperglycemia. The liver is also central to the control, storage and distribution of fats. Apolipoproteins synthesized in the liver are used to assemble and distribute lipoproteins containing triglycerides and cholesterol esters to other tissues in the form of VLDL, which is degraded to form LDL. Cholesterol is recovered by the liver through LDL receptors. The liver also receives cholesterol from other tissues in the form of HDL by way of HDL receptors. This process is under complex hormonal and dietary control. Corticosteroids promote the distribution of lipids and reduce the uptake of cholesterol by the liver causing dyslipidemia and ultimately atherosclerosis. The influences of corticosteroids are both direct and indirect by way of glucocorticosteroid receptor (GR) binding altering the expression of other transcription factors such as sterol regulatory element-binding proteins (SREBP-1) which may in turn regulate other genes.

Previously, we treated a group of adrenalectomized, adult male rats with a single bolus dose of the synthetic glucocorticoid methylprednisolone (MPL). A control group and treated animals were sacrificed at 16 time points over a 72 hour period following dosing. A variety of

individual measurements were made on the livers from these animals and the results have been used to construct pharmacokinetic/pharmacodynamic (PK/PD) models (Jin et al., 2003). Considering the broad impact of corticosteroids on the liver, the approach of measuring changes in individual genes and biomarkers only provides a limited view of the system. To obtain a global picture of the hepatic response dynamics to MPL, mRNA from these livers were applied to individual Affymetrix gene chips. The result was six clusters with very high within class correlation containing 143 unique genes and mechanism-based PK/PD models for each of the six clusters was proposed (Jin et al., 2003). However, this effort required both the identification of the number of clusters as well as the manual identification of the genes which were hypothesized to be important.

We propose an algorithm seeking the global genomic response of the liver to MPL and develop a feature-based gene selection method which attempts to detect salient features and global shape characteristics of the expression profiles. A key motivating arguments for this method is the realization that in the presence of noise and uncertainties associated with measuring mRNA abundance, looking for specific quantifiable metrics may not necessarily yield the most informative interpretation (Tilstone, 2003). However, most robust, coherent and dominating qualitative features and similarities are a more informative proxy for the information content of the expression experiment.

The raw data is initially transformed into sequences of symbols which are further analyzed for consistencies. This algorithm works off the assumption that genes which are relevant to the underlying dynamics of the system have two essential characteristics. Firte, they are part of a concerted mechanism and should possess expression profiles which are temporally

consistent with the expression profiles of other genes involved in related processes. Second, informative genes ought to contribute to global deviations away from the baseline state. Therefore, the algorithm performs a fine-grained clustering which results in hundreds of clusters. We then evaluate the ability of each of these individual clusters to satisfy these two constraints, thereby linking the selection process with the clustering result. Our underlying assumption is that hidden in the temporal microarray expression data is a reduced set of transcriptional signatures that have captured the essential dynamics of the cellular response. We are not merely interested in clustering all expression responses, but rather in identifying the subset of elementary responses that capture this intrinsic dynamic response of the system. We propose a quantification of the intrinsic response, through our definition of the transcriptional state, and chose among the multitude of micro-clusters the subset that maximizes deviations from homeostasis. Once the intrinsic dynamics has been revealed, it can be used for the development of PK/PD models. Current algorithms focus on clustering all responses and thus do not allow for qualitative interpretations and assignment of significance to specific subsets. As such, our approach is unique and best suited for the specific task at hand.

Methods

Experimental Design

Liver samples were obtained in a previously performed animal study (Sun et al., 1999). All procedures involving experimental animals adhered to the “Principles of laboratory Animal Care” (NIH publication 85-23, 1985) and were reviewed by our institution’s animal care and use committee. Male adrenalectomized (ADX) Wistar rats (Rattus rattus) weighing 225–250 g were obtained from Harlan Sprague-Dawley (Indianapolis, IN.). The control rats have had their

adrenal glands removed. This removes all corticosteroid mediated circadian variance in the data. Furthermore, given that the administration of corticosteroids represents a large perturbation to the system as measured by the mRNA, we feel that the circadian impact on the CS response is minimal (Oishi et al., 2005).Animals were allowed free access to rat chow (Agway, RMH 1000) and 0.9% NaCl drinking water. They were housed in a room with a 12 hr light/12 hr dark cycle, a constant temperature of 22°C and were allowed to acclimatize to this environment for at least 1 week. One day prior to the study, all rats were subjected to right external jugular vein cannulation under light ether anesthesia. Cannula patency was maintained with sterile 0.9% NaCl solution. Four animals were designated as controls and were cannulated but only received vehicle. All remaining animals received a single 50 mg/kg dose of methylprednisolone sodium succinate (Pharmacia-Upjohn Company, Kalamazoo, MI) via the cannula over 30 s. Rats (3/time) were sacrificed by exsanguination under anesthesia at 0.25, 0.5, 0.75, 1, 2, 4, 5, 5.5, 6, 7, 8, 12, 18, 30, 48, and 72 hr after dosing. The four control rats were designated as time 0. The sampling time points were selected based on previous studies describing GR dynamics and enzyme induction in liver and skeletal muscle. Livers were rapidly excised, flash-frozen in liquid nitrogen, and stored at -80°C. Frozen tissues were ground into powder using a liquid nitrogen chilled mortar and pestle.

Microarrays

Liver powder (100 mg) from each rat was added to 1 ml of pre-chilled Trizol® Reagent (InVitrogen Carlsbad, CA). Total RNA extractions were carried out according to manufacturer's directions. Extracted RNAs were further purified by passage through RNeasy mini-columns (QIAGEN, Valencia, CA) according to manufacturer's protocols. Final RNA preparations were resuspended in nuclease-free water and stored at -80°C. RNAs were quantified

spectrophotometrically.; purity and integrity were assessed by agarose gel electrophoresis. All RNA samples exhibited 260/280 ratios between 1.8 and 2.0, and exhibited discrete ribosomal bands on agarose formaldehyde gels, indicating minimal sample degradation. The biotinylated cRNAs were hybridized to 47 individual Affymetrix GeneChips® Rat Genome U34A (Affymetrix, Inc., Santa Clara, CA), which contained 8799 probe sets. This entire data set has been submitted to the NCBI Gene Expression Omnibus database (GDS253) and is also available on line at http://pepr.cnmcresearch.org.

Temporal Analysis of Gene Expression Data

Because the ADX animals lack endogenous corticosterone, MPL represents a stimulus that perturbs the balance of the system and the time series design allows us to evaluate deviations from baseline and its return to the original state. Global analysis of the dynamics of the system involves two critical steps. The first is the identification of major expression patterns. These are temporal subpatterns that are associated with a set of genes and are maximally different than all other subpatterns in the time series. The second is the characterization of the transcriptional dynamics of the system.

Selection of Informative Genes

A critical step in processing high throughput gene expression data is the selection of genes for further analysis. Consideration of individual variables such as measurements of the expression of TAT mRNA and protein derives from the standard hypothesis-driven approach. Given the overall structure of microarray data, the analysis problem is entirely different. Although one might expect that a particular gene has importance and look at its data, the real challenge is to globally identify the fraction of genes that are relevant to the response of the system.

Most of the current methods for the selection of relevant gene expression profiles rely upon statistically significant changes in expression level. The most simple and probably the most commonly used technique is the n-fold test. In our previous analysis, the genes were filtered with the n-fold algorithm where genes which were up/down-regulated by 1.5 times for 4 time points or more were chosen as significant(Almon et al., 2007). Other techniques utilize statistical methods such as the t-test, ANOVA and SAM (Millenaar et al., 2006). These tests essentially look for statistically significant changes in the expression data from the baseline. This allows for a differentiation between activated/non-activated states, but does not isolate genes which show coordinated changes in their responses over time.

More recently, gene expression profiles have been selected via the over-representation of a particular shape in the expression profile. Techniques such as SLINGSHOTS (Yang et al., 2007), STEM (Ernst and Bar-Joseph, 2006), and QT-Clustering (Heyer et al., 1999) fall under this category. These techniques seek dominant patterns in the data, ascertain which genes correspond, and select them as biologically significant. These methods work under the notion that large groups of co-expressed genes tend to be more significant than genes that do not show such a high degree of co-expression. Specifically for this analysis the SLINGSHOTS algorithm was selected because it not only identifies biologically significant genes, but also it has the ability to identify the global response characteristic along with the corresponding genes.

The SLINGSHOTS algorithm is broken up into the following steps:

1. Identification of over-represented populations.
2. Selection of over-populated clusters that reflect some unknown global dynamic.

The most common methods for assessing whether a given expression profile is associated with a large number of co-expressed genes are the various clustering techniques. For the

purposes of SLINGSHOTS, any micro-clustering technique can be utilized, defined as any clustering method which utilizes a large number of clusters. With a large number of small clusters, it becomes easy to assess the density or the over-representation of a given expression profile or pattern. We have elected to utilize a hash based clustering technique first proposed by Lin et al (Lin et al., 2003) because it is a deterministic and efficient method for identifying over-represented clusters. The clusters identified have the same minimum correlation thereby allowing us to easily isolate the over-populated patterns.

Identification of major expression patterns

The hashing based clustering consists of the following steps:

1. Z-score normalization of the signal.
2. Piecewise averaging of adjacent points.
3. Conversion into symbolic representation.
4. Conversion of the symbolic representation into an integer, and genes hashing to the same integer are treated as belonging to the same cluster.

The z-score normalization, given in **Eq.(1)** where μ is the mean of the signal and σ is the standard deviation of the signal, normalizes the expression profiles (y_t) to vary around a mean of zero and a standard deviation of 1.0.

$$\hat{y}_t = \frac{y_t - \mu}{\sigma} \qquad (1)$$

This allows metrics such as the Euclidian distance to return the same result as the Pearson correlation.

The piecewise averaging signal essentially smoothes and shortens the signal. This is one of the two parameters which must be determined beforehand. In general, if a dataset is shorter than 9 time points, no averaging needs to be conducted, whereas if the dataset is longer, the

JPET Fast Forward. Published on December 17, 2007 as DOI: 10.1124/jpet.107.133074
This article has not been copyedited and formatted. The final version may differ from this version.

smallest number of adjacent points should be averaged, thereby maintaining a total length of around 9. This removes high frequency components in a similar fashion to a low-pass filter as well as maintaining numerical tractability. The latter issue comes into play during the fourth step, but essentially, this hashing method involves an exponential expansion of the hash space in relation to the signal length. This piecewise averaging is conducted as per **Eq.(2)** where**,** *T* represents the overall length of the signal, *w* represents the new desired length of the signal, and *k* is a free variable that ranges from 1 to *w*.

$$y_k = \frac{T}{w} \sum_{t=(k-1)\frac{T}{w}+1}^{k\frac{T}{w}} Y_t \qquad (2)$$

The third step involves converting this piecewise average into a series of symbols. This is accomplished via the use of Gaussian breakpoints as illustrated in **Figure 1.** While any method for discretizing the signal can be used, Gaussian breakpoints assigns a symbol with equal probability given randomly generated data. Given the latter, the distribution of hash values ought to correspond to that of an exponential distribution (Indyk et al., 1997). This property allows for the assessment as to whether the dataset itself shows significant coordination among genes. It also offers guidance as to the number of breakpoints needed which should be chosen so there is a large deviation from the exponential distribution. For this dataset with seventeen time points, we chose to piecewise averaging of every 2 points and have 3 breakpoints.

The breakpoints can be obtained from tables of the Gaussian CDF, or via **Eq.(3)** by solving for *x* where *z* is the number of breakpoints, and *n* is an index variable in the range of 1 to *z*, and erf is the error function.

$$\frac{n}{z-1} = \frac{1}{2}\left[1 + erf\left(\frac{x}{\sqrt{2}}\right)\right] \qquad (3)$$

The overall process of converting the signal into its symbolic representation is shown in **Figure 1.**

The final part of the process involves converting the symbolic representation into an integer. This is accomplished by treating the sequence as a base *N* number where *N* is the number of breakpoints. By then converting to a base 10 number as in **Eq.(4)** we obtain an integer *(h)* where, if two genes hash to the same integer, they have similar expression profiles, where c is symbol obtained from **Eq.(3)**, w is the length of the sequence, and a is the size of the alphabet.

$$h = 1 + \sum_{j=1}^{w} \left[ord(c_j) - 1 \right] a^{w-j} \qquad (4)$$

Characterization of the transcriptional state of the system and extraction of the most informative expression patterns

After the conversion of the gene expression pattern into a set of motifs (hash values), we then evaluated which of these motifs were required to capture the non-random progression of the entire expression profile over time. This was done by first defining a concept we term the transcriptional state. This allows for the construction of a metric that characterizes the inherent dynamic state of the system that allows monitoring of the dynamic progression and evolution of the system. Such a quantifiable metric allows us to evaluate just how informative the selected subset of probe sets is. The basic premise is that there exists a core set of genes whose transcriptional machinery is most affected by the stimulus. Furthermore, this core set of genes represents the fundamental response of the system and thus accounts for its essential dynamics. In order to characterize the dynamic state of the system we treat the expression levels of each probe set as variables that follow a specific, albeit unknown, distribution. The drug, will alter these distributions over time to perturb the underlying transcriptional machinery of each gene, quantified by the corresponding amounts of mRNA. Given such perturbations, it would be

expected that, over the course of time, the distribution of expression values of the most informative subsets of genes would show greater deviations from both the Gaussian and the baseline (control) expression levels.

To quantify this observation, the Kolmogorov-Smirnov (K-S) test, is employed. The K-S test is applicable to unbinned distributions that are functions of a single independent variable. The list of data points over time can then be easily converted to an unbiased estimator of the cumulative distribution function of the probability distribution from which the expression values were drawn. Therefore, truly informative subsets of genes are the ones that have the ability to capture significant deviations from the base distribution. The K-S test is a simple yet effective way of comparing two distributions and has had many applications such as the estimation of diversity in chemical libraries (Rassokhin and Agrafiotis, 2000).

The K-S is defined as the maximum absolute difference between two cumulative distribution functions. For each time point, we estimate a cumulative distribution function (CDF) of the expression values, after a double normalization step. The first normalization of the expression profiles sets the range of all profiles to the same scale. The second normalization is performed so that distributions of the same family are lumped together whereas distributions of different families are quantified as different. For instance, this will lump all normal distributions no matter the parameter together whereas a normal parameter and an exponential parameter would be classified as different. By doing so, we can ascertain whether different mechanistic factors are affecting the distribution of gene expressions at the different time points.

The base distribution is the corresponding CDF prior to administration of the drug. The K-S statistic *(D)* is this defined as were n is the number of genes in a given population and *i* is the index of a particular gene and its associated position on the CDF:

JPET Fast Forward. Published on December 17, 2007 as DOI: 10.1124/jpet.107.133074
This article has not been copyedited and formatted. The final version may differ from this version.

$$D = \max_{1 \le i \le n} \left| F(Y_i) - F(Y_i(0)) \right| \quad (5)$$

where $F(Y_i(0))$ is the cumulative distribution of the expression values at time $t = 0$. This statistic allows a metric that defines the magnitude of the difference between two distributions to be computed. Since the data is presented as a time series, at each time point a value for the K-S statistic is obtained. Therefore, the overall metric becomes:

$$D = \max_{t} \max_{1 \le i \le n} \left| F[Y_i(t)] - F[Y_i(0)] \right| \quad (6)$$

The application of the K-S test over time allows us to quantify just how much the CDF of a particular sub-set of genes deviates from the corresponding CDF at time $t = 0$ (control). The most sensitive subset exhibits the largest deviations from the control. Once the subset is specified then it can be characterized based on its corresponding D value. The individual genes in a subset are then defined by the previously described hashing procedure. We have implemented a simple greedy algorithm that selects peaks based on their ability to maximize the deviation from the control distribution of expression values. The basic steps of the algorithm are as follows:

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

(vi) Evaluate $D(k) = \max_{t} \max_{g_i \in \Sigma} \left| F[Y_{g_i}(t)] - F[Y_{g_i}(0)] \right|$

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

The iteration count *k* corresponds to the number of peaks that are incorporated at each step, *S(k)* is the set of hash values that have been considered up to iteration *k*, *N(h)* is the number of genes probe sets that have been assigned to a particular hash value *h*, *h** is the motif values that is most populated at each iteration, *G(k)* is the subset of genes g_i, that have hashed to *h*, while *S* is the cumulative set of genes included at each iteration. *D(k)* is the K-S statistic evaluated at iteration *k* and is calculated using the set *S* of genes. Once a peak and its corresponding *S* probe sets, have been included, then the corresponding hash value is eliminated so that it is not considered again. The search is performed in the space of peaks, as opposed to individual probe sets, and peaks (along with the corresponding probe sets) are added provided that a clear deviation from the control state is observed.

Results

Major Expression Patterns

Figure 2 depicts the distribution of motif values for rat liver gene expression after dosing with MPL. The algorithm isolated 529 out of a total of 8799 probes into 12 clusters. The justification for this selection is given in **Figure 3**. The first 12 clusters, produced the most evident deviation from baseline of the transcriptional state. While there exist other overpopulated clusters which are not included,. we do not dispute the fact that these genes may be significant, nor claim that the selected genes are more biologically significant. Instead, what we claim is that given factors such as noise, the selected genes are the ones from which the global dynamics are

most visible. This is significant because it is from the global dynamics as PK/PD response model can be constructed. We utilize the Signal to Noise Ratio (SNR) defined as $SNR = 20 * \log_{10} \frac{\mu(g)}{\sigma(g)}$ as a measure of signal quality. The SNR measures signal quality by comparing the scale of the mean vs. the scale of the standard deviation. The μ represents the mean of the signal and σ represents the standard deviation of this reported value (Greshock et al., 2007). The value is reported in (decibels) dB which is the log transformation of the value. The greater the mean is in relation to the variance the greater the SNR. Given the logarithmic formulation of the SNR, a signal in which the standard deviation was the same as the mean would have an SNR of 0, whereas if the standard deviation was smaller than the mean, the SNR would be positive, and negative otherwise. The genes selected via the algorithm have a minimal SNR of 2dB and a mean SNR of 16dB with the majority of the genes (67%) having an SNR of above 12 dB. This means that in all cases, the signals selected, the mean values were significantly greater than the variance. SNR was chosen as a method for quantifying quality because of it offers a rough measure as to how effective signal processing methods are at extracting the intrinsic dynamics.

Figure 4 provides the 12 expression versus time profiles for all of the genes in each cluster. They exhibit clear characteristics of early up- or down-regulation and the expression levels in these patterns return to baseline with time. It is evident that the hash-based clustering has yielded groups of genes with highly correlated activity. The selection of these clusters does not preclude other genes and clusters from having a biological role. Rather, these 12 clusters represent the majority of responses that comprise what we deem to be the intrinsic system dynamics.

Characterization of the transcriptional state of the system

Figure 5 depicts the deviation from the t = 0 distribution for the informative probe sets belonging to the 12 primary motifs selected by the algorithm. The temporal evolution of the transcriptional state as well as the plot of the objective function over time (**Figure 6**) shows that the initial deviation is followed by an eventual return to the initial state. These 12 peaks reveal the minimum number of expression signatures (not individual genes) whose presence is critical for reproducing the dominant transcriptional response of the system. The K-S statistic reflects the overall dynamics with a large deviation from the baseline before hour 10, and then a return to the initial state. This is in agreement with an acute pharmacodynamic response in which the responses to the single bolus dose of drug occur early and then are reduced over time as the drug is cleared from the system.

Dynamic Response Model

Aside from being useful as a metric for the selection of informative genes, the K-S statistic itself offers valuable information as to the overall response of the system to the single bolus dose of MPL. Utilizing the K-S statistic rather than individual genes for model creation allows us to treat the system as an aggregate and obtain a time constant of drug activity rather than time constants of specific genes. This negates the need to first identify a candidate gene as was done in the original approach. From the KS plot in **Figure 6**, we hypothesized that the overall drug response could be modeled as a closed loop LTI (Linear Time Invariant) model. This does not mean that the underlying corticosteroid response is linear, just that for this single drug dose, the system can be approximated as a linear mass-action system. We acknowledge the fact that there exist significant non-linearities in the system as evidenced by the fact that the overall corticosteroid response shows a tolerance mechanism during repeated dosing(Sun et al.,

1998). This violates the linearity properties of the LTI model. This simplification was based upon the fact that the overall shape of the damped response was very similar to the original response. This means that the response can be modeled either as a nonlinear system or a piecewise linear system by taking the non-linear portions of the model and piecewise linearizing them. In order to obtain the primary components of the system, we have elected to do the latter. This simplification is then used to provide specific intuitions about the response of the system. This model can later be modified to include specific non-linear components such as receptor saturation and receptor ligand interactions which would then allow us to model behavior such as the dose dependent nonlinear response as well as the aspect of tolerance. This approach is something dissimilar to the mechanistic based modeling previously proposed in which intuitions about the system is used to create a comprehensive model(Jin et al., 2003), in that we seek to create a framework response, then incorporate additional factors such as nonlinearities when faced with additional dat. The additional of these additional factors essentially allow us to hypothesize the behavior of important control elements in the system, rather than requiring their knowledge *a priori* and represents the difference between a hypothesis driven approach vs. a data driven discovery approach.

A linear time invariant (LTI) model satisfies two properties. First, it has a non-time dependent response. This means that there is no implicit time definition and so the modeling equation as defined as $f(x(t))$, where $x(t)$ is the input, rather than $f(x(t), t)$. This property is supported by use of ADX rats to eliminate the circadian effects of endogenous corticosterone. The second is that the system must reflect the principle of superposition. This means that $f(x(t) + y(t)) = f(x(t)) + f(y(t))$, in which both $x(t)$ and $y(t)$ are separate inputs. The primary attraction of this method is that it greatly simplifies the overall construction of the model.

The LTI formulation allows for the use of the convolution integral to determine the response of the system from an input **Eq.(7)**. The use of the convolution integral is attractive because it allows the system to be described in the Laplace domain **Eq.(8)**, which allows for identification of mathematical representations of physical factors or elements which amplify or dampen the input response *f(x)*.

$$f * g = \int_0^t f(\tau) g(t-\tau) d\tau \tag{7}$$

The gross schematic of a closed loop LTI model is shown in **Figure 7**. The model consists of a gain term *k(s)* and a feedback term *g(s)*. The feedback term relates the current state of the system to a future state, and does not strictly imply the presence of the gene product activating a pathway which will then deactivate it. The simplest mechanism that could function as this feedback term would be a reservoir which couples the synthesis and degradation rates of mRNA.

Nominally, an LTI system is a set of differential equations (Oppenheim et al., 1997). In the simplest case, a single input/single output model (SISO), the differential equation can be written as in **Eq.(8)** where *(m)* and *(n)* represent the degree of differentiation (first, second, third derivative).

$$\sum_{n=0}^{p} a_n y^{(n)} = \sum_{m=0}^{q} b_m x^{(m)} \tag{8}$$

The Laplace transformation is defined in **Eq.(9)**. It is closely related to a Fourier transform and is often used in electrical engineering for systems identification (Franklin et al., 2002).

$$L[f(t)](s) = \int_0^\infty f(t) e^{-st} dt \tag{9}$$

The advantages of the Laplace representations are two-fold. The prediction of the response to other inputs such as chronic drug infusion can be achieved through simple algebraic manipulation. Additionally, the Laplace domain differential equation has distinct physical meaning and can give insights as to the underlying mechanisms that govern the observed response.

The conversion of a first-order differential equation into the Laplace domain is given in **Eq.(10)**.

$$y' + ay = x$$

$$L[y' + ay] = L[x] \tag{10}$$

$$Y(s) = \frac{1}{(s+a)} X(s)$$

This generalized form has the corresponding Laplace transform given in **Eq.(11)**, which in our case has the constraint where $m > n$.

$$Y(s) = \frac{\prod(s + a_n)}{\prod(s + b_m)} \tag{11}$$

This additional constraint imposes stability upon the system which is not universally required, but in the context of a biological responses to corticosteroids is reasonable. An unstable response would indicate a changing of state after the drug has been cleared from the system. This could result in death as the organism would be unable to recover homeostasis. In **Eq.(11)** we can represent the system as the quotient of two polynomials. The polynomial is then factored with the terms in the numerator representing the "zeros" of the equation and the terms in the denominator representing the "poles" of the equation. The factorized terms in the numerators function as inductors while in the denominator function as capacitors. The biological analog of a capacitor is a compartment which accumulates the buildup of the input signal or mRNA, and the

biological analog of an inductor is the inertia of the signal. For instance in the circulatory system the mass of the fluid being displaced can act as an inductive element (Ferrari et al., 2003).

The feedback as described in this system does not necessarily correspond to the biological notion of feedback. It only requires that the rate of change, i.e. *dx/dt*, be dependent upon *X* or the amount already present in the system. This can be handled via the more familiar indirect response models where the amount of the signal or mRNA directly affects the sensitivity to the signal or the production of mRNA. However, it could also be modeled as a mass action system in which the rate of loss is dependent upon concentrations already present through changes in either degradation or production rates. However, at this point, we do not consider the mechanism for feedback, but only establish that there is one.

Part of the reason behind SLINGSHOTS was the extraction of the global dynamic. **Figure 7** integrates the expression profiles of different genes. Utilizing the Laplace formulation, we fitted the global dynamic to the general form given in **Eq.(11)**, while minimizing the overall number of terms. This was accomplished via the LSIM command in Matlab which takes the Laplace representation and is able to simulate an impulse response and leads to the response observed in **Figure 7**. It was found that the data could have been fitted via **Eq.(12)**, an equation with no terms in the numerator except for a constant scaling factor which can be thought of as a constant amplification factor of the mRNA signal, and two terms in the denominator indicates the need for two capacitance elements.

$$Y(s) = \frac{k}{(s + a_0)(s + a_1)} \qquad (12)$$

The first capacitance element (a_0) is the circulation which acts as a compartment which can store a portion of corticosteroid dose. The second capacitance element (a_1) are the cells where the

signal alters the production of mRNA thereby having an indirect rather than direct effect. This reflects the original hypothesis that the activity of corticosteroids is mediated via drug from plasma interacting with tissue receptors.

Given the agreement between this model and those previously derived (Dayneka et al., 1993), we believe that the global dynamic obtained via our K-S statistic acts as a good surrogate for the global activity of a population of genes, thereby obviating the need to specifically identify genes that respond to corticosteroids through *a priori* knowledge. The primary advantage of utilizing both the K-S statistic as well as the pole-placement model for modeling identification is that it is a data driven approach, and is independent of researcher bias. It does this without any loss in generality as seen in the agreement between the gross mechanistic aspects of the indirect response models (Dayneka et al., 1993) and the mechanical underpinnings of our model.

Functional Dynamics

To examine the functional dynamics of the system we conducted an extensive literature review of all affected genes (On line supplementary Tables 1-14). The genes were separated into different functional categories that spanned multiple clusters and the hierarchical clustering and visualization gene tree approach (Eisen et al., 1998) was applied as modified by the GeneSpring software. We used this algorithm to construct a dendrogram of genes with similar patterns based on the Pearson correlations. A negative aspect of this tool is the assumption that the points in the time series are equally spaced. Notwithstanding this drawback, gene trees provide an excellent method of visualizing the dataset. It was necessary to first transform the data so that the values for all probe sets were within the same range. Values for each probe set on each chip were expressed as a ratio of the mean of the four control values for that gene, which we refer to as “normalized intensity”. Thus the average of each probe set has a value of 1.0 at zero time and

either increases, decreases, or remains unchanged relative to controls over the time series. Yellow in the graph represents an expression ratio around 1, or no change. The color progressing toward red indicates a normalized value greater than 1, or up-regulation, and the color toward blue indicates a value less than 1, or down-regulation from control levels. **Figures 8a-l** provide the dendograms for the twelve functional categories ('Other' and EST were excluded). These figures provide an overview of the effects of MPL on the overall functional dynamics of the liver. For example, the dominant red color in **Figure 8a** demonstrates that the general effect on transcription and translation is to enhance the expression of the particular genes. Interestingly, there is limited down-regulation of genes seen as a blue band at the bottom. Notable amongst these genes is RXR, which is strongly down-regulated. This likely is related to alteration of lipid metabolism seen in **Figure 8k**. In contrast, **Figure 8b** shows that there is both significant enhancement and down-regulation of genes involved in signaling. **Figure 8c** shows that the dominant effect of MPL is down-regulation of genes involved in small molecule metabolism. The interesting exceptions of glutamine synthestase and ornithine decarboxylase in cluster 1 along with argininosuccinate lyase and carbonic anhydrase in cluster 4 are involved in disposal of the ammonia produced by gluconeogenesis from amino acid carbon. These figures also demonstrate that the majority of the effects of MPL are finished by12 hours, although some effects on immune related genes, **Figure 8e** and mitochondrial genes, **Figure 8j** persist well beyond this time.

Discussion

The primary mechanism of action of corticosteroids is alteration of the expression of genes. In contrast to drugs that have direct actions where there is a relationship between the drug concentration at the effect site and magnitude of the response, corticosteroids have indirect effects on gene turnover that persist long after the drug has dissipated. Most of the extracted profiles have similar dynamics as those found in the previous analysis which consists of an initial deviation from the baseline followed by a relaxation back to the initial state(Jin et al., 2003).

This experiment was designed such that the single dose of MPL initiated molecular events in time that continued until the complex system returned to its initial equilibrium state. The pattern recognition approach within the context of the rich time series design allowed us to observe the overall dynamics of the system as it is pertrurbed and re-equilibrates. Our method identifies the dominant motifs of this dynamic process, and the genes that comprise these profiles. This approach is different from the commonly used "clustering algorithms" such as K-means or SOM where some similarity measurement (eg. correlation or Euclidian distance) is used to collect genes into a predetermined number of similarity groups. The present approach seeks in an unsupervised, "top down" manner to identify the dominant motifs governing the dynamics of the system. The use of this system dynamics approach allowed us to both characterize the behavior of this complex system and to parse into groups the contributing elements of the system.

One of the difficulties with creating a PK/PD model from individual genes or clusters is that it is difficult to tease out the interplay between the different clusters. Ideally, representative genes would respond only to the input of corticosteroids. This is problematic for two reasons. First, it is difficult to determine whether the gene responds only to MPL. Methods such as

computational prediction of transcription factor binding sites are not sufficiently sensitive to fully identify the possible regulators of a given gene, and even techniques such as Chip-Chip experiments cannot identify which of the possible regulators are active in a given study. Second, the derivative effects of corticosteroids such as the elevation of circulating glucose may have effects on the overall system, and such secondary changes may not be captured by a single gene.

Previously we generalized a representative gene to identify genes with similar patterns(Almon et al., 2007).Models were developed describing responses of these sets of genes (Jin et al., 2003). This approach allowed for the incorporation of secondary effects as well as the identification of time-lagged responses, but was not able to account for complexities such as the possibility that each of the genes may interact with each other. Our previous corticosteroid models were simplified as one dose of MPL was given and different interactions were not considered. The models generated in such a fashion were not always applicable for a chronic infusion regimen (Almon et al., 2007). Attempts have been made to add back into the model the links in the form of other biosignals and transduction steps and further work will be needed to identify and quantify the transcriptional links that tie together the disparate subsystems.

Creating models from the overall dynamics of the system represents a level of abstraction which bridges the gap between the simplicity of a single gene model, and more complete model(s) derived from multiple genes. It allows for the creation of models which do not require the explicit identification of interactions between different genes while still exhibiting the salient properties of the system. In this case, we posited a central necessity of having a two-component model which agrees with the general properties found in the indirect response models (Dayneka et al., 1993). Building a model from the global dynamics helps provide initial intuitions about the system. It is more complete than single gene models because it allows for a composite response

which can come from multiple genes. It lumps the different expression profiles into a single dynamic. Additionally, the global response provides an alternative but complimentary view of the response dynamics to drug administration rather than looking at individual genes or systems which are affected. Instead of considering the individual gene time courses, one considers the degree, the extent, and how long the drug has a tangible effect upon the overall system. This representation provides an overall context for the development of more complex models for those genes of special interest.

The intersection of the 216 probes extracted in the original analysis and the 529 probes extracted via the new method, there was an interaction of only 56 probe sets. While this may seem like a small amount, utilizing ANOVA on the dataset with a $p < 1.13x10^{-4}$ yielded 267 genes with an intersection of only 36 genes. A large majority of the genes that intersected between the two datasets were found in Cluster 1 and 6 in the original analysis with and a single gene found in Cluster 4. These correspond to the up-regulated, down-regulated cluster, and one which was originally termed “biphasic” in which there was an initial down-regulation and then an up-regulation above the initial baseline level before returning to the final steady state. In this analysis we will not be differentiating the biphasic response as a separate mechanism because it can be reconstructed via the standard 2-pole LTI model as easily as the other systematic profiles and suggests the existence of a complex pole. Physically, the LTI model can be implemented as a delay in the system either through inertial means such as differences in diffusion rates or in the case of the CS models, the implementation of a lag term in which there is an intermediate genes which is transcribed which later activates another gene as denoted by BS in the original CS model(Jin et al., 2003).

Two of the clusters which were not found in our analysis correspond to clusters 2 and 5 which are sparsely populated and therefore do not seem to be part of any significant coordinated event. Of greater concern however is cluster 3 in the original analysis, which again shows a systematic overshoot profile but was not selected via our SLINGSHOTS algorithm. This may be due to the greater variability in the cluster due perhaps to intermediate signals which are not consistent between the genes as evidenced by the large difference in the k_{d_BS} in the genes of that cluster when associated with the fifth generation model, showing that they have different intermediate signals, and therefore are not direct effects of corticosteroid on mRNA expression levels but rather secondary effects which may still be important. The breakdown of cluster coherence as the corticosteroid response progresses through multiple layers of the cascade is something which can be addressed with parameters with less granularity such as increasing the windows size from 2 time points to 3 time points. However, this loss of coherence weakens the concept of co-expression implying co-regulation(Wolfe et al., 2005). Therefore, we believe that the SLINGSHOTS algorithm represents one of the many steps needed to fully decipher complex transcriptional mechanisms and therefore does not negate the value of the previous analysis.

The original motivation for re-examining the extensive gene array data was to determine whether a more comprehensive set of genes affected by MPL could be selected. In this formulation, we identified 529 probes as opposed to the 192 which had been found initially. The high correlation of the genes selected lends confidence in the clusters. From these genes, we have isolated many major components which are responsible for the effects of corticosteroids such as transcriptional signaling, small molecule metabolism, as well as genes responsible for controlling metabolic shift changes.

Most importantly, we have identified the global dynamics of the system and proposed a relatively simple model which can be used as the basis to model the dominant PK/PD responses of the liver to MPL. While this model is less complete than the previously devised PK/PD models, it has the primary response feature of the model, namely turnover of mRNA. While the interplay between the different genes has not been identified or quantified, the current model explains the general response pattern. The identification and quantification of the systems which are responding as part of this global response will be a logical next step. In essence, we have created a framework within which additional PK/PD models for drug responses can be developed. The overall value of this algorithm is not so much its ability to select genes, but rather its selection of elementary response profiles as well as describe the global response dynamics of the system.

References

Almon RR, Dubois DC and Jusko WJ (2007) A microarray analysis of the temporal response of liver to methylprednisolone: a comparative analysis of two dosing regimens. *Endocrinology* **148**:2209-2225.

Chikanza IC (2002) Mechanisms of corticosteroid resistance in rheumatoid arthritis: a putative role for the corticosteroid receptor beta isoform. *Ann N Y Acad Sci* **966**:39-48.

Cronstein BN, Kimmel SC, Levin RI, Martiniuk F and Weissmann G (1992) A mechanism for the antiinflammatory effects of corticosteroids: the glucocorticoid receptor regulates leukocyte adhesion to endothelial cells and expression of endothelial-leukocyte adhesion molecule 1 and intercellular adhesion molecule 1. *Proc Natl Acad Sci U S A* **89**:9991-9995.

Dayneka NL, Garg V and Jusko WJ (1993) Comparison of four basic models of indirect pharmacodynamic responses. *J Pharmacokinet Biopharm* **21**:457-478.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**:14863-14868.

Ernst J and Bar-Joseph Z (2006) STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* **7**:191.

Ferrari G, Kozarski M, De Lazzari C, Gorczynska K, Mimmo R, Guaragno M, Tosti G and Darowski M (2003) Modelling of cardiovascular system: development of a hybrid (numerical-physical) model. *Int J Artif Organs* **26**:1104-1114.

Franklin GF, Powell JD and Emami-Naeini A (2002) *Feedback control of dynamic systems*. Prentice Hall, Upper Saddle River, NJ.

Greshock J, Feng B, Nogueira C, Ivanova E, Perna I, Nathanson K, Protopopov A, Weber BL and Chin L (2007) A comparison of DNA copy number profiling platforms. *Cancer Res* **67**:10173-10180.

Heyer LJ, Kruglyak S and Yooseph S (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* **9**:1106-1115.

Indyk P, Motwani R, Raghavan P and Vempala S (1997) Locality-preserving hashing in multidimensional spaces, in *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing* pp 618-625, El Paso, Texas.

Jin JY, Almon RR, DuBois DC and Jusko WJ (2003) Modeling of corticosteroid pharmacogenomics in rat liver using gene microarrays. *J Pharmacol Exp Ther* **307**:93-109.

Jusko WJ JJ, Dubois DC, Almon RR (2005) Sixth-Generation Model for Corticosteroid Pharmacodynamics: Multi-Hormonal Regulation of Tyrosine Aminotransferase in Rat Liver *J. Pharmacokin. Pharmacodyn.;* .

Lin J, Keogh E, Lonardi S and Chiu B (2003) A symbolic Representation of Time series, with Implication for Streaming Algorithms, in *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery`*, ACM, San Diego, CA.

Millenaar FF, Okyere J, May ST, van Zanten M, Voesenek LA and Peeters AJ (2006) How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC Bioinformatics* **7**:137.

Oishi K, Amagai N, Shirai H, Kadota K, Ohkura N and Ishida N (2005) Genome-wide expression analysis reveals 100 adrenal gland-dependent circadian genes in the mouse liver. *DNA Res* **12**:191-202.

Oppenheim AV, Willsky AS and Nawab SH (1997) *Signals & systems*. Prentice Hall, Upper Saddle River, N.J.

Rassokhin DN and Agrafiotis DK (2000) Kolmogorov-Smirnov statistic and its application in library design. *J Mol Graph Model* **18**:368-382.

Sun YN, DuBois DC, Almon RR, Pyszczynski NA and Jusko WJ (1998) 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* **26**:619-648.

Sun YN, McKay LI, DuBois DC, Jusko WJ and Almon RR (1999) Pharmacokinetic/Pharmacodynamic models for corticosteroid receptor down-regulation and glutamine synthetase induction in rat skeletal muscle by a Receptor/Gene-mediated mechanism. *J Pharmacol Exp Ther* **288**:720-728.

Tilstone C (2003) DNA microarrays: vital statistics. *Nature* **424**:610-612.

Wolfe CJ, Kohane IS and Butte AJ (2005) Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics* **6**:227.

Yang E, Maguire T, Yarmush ML, Berthiaume F and Androulakis IP (2007) Bioinformatics analysis of the early inflammatory response in a rat thermal injury model. *BMC Bioinformatics* **8**:10.

Footnotes

EY and IPA acknowledge support from NSF grant 0519563 and the EPA grant GAD R 832721-010. RRA, DCD and WJJ acknowledge support from NIH grant GM 24211 from the National Institutes of General Medical Sciences.

Legends for Figures

Figure 1: A schematic of converting a expression profile into a hash value

Figure 2: The motif distribution of our data. Motifs 1-12 were selected as relevant motifs

Figure 3: After the critical number of clusters have been added. While there may be information present in additional cluster, it degrades the overall

Figure 4: The z-score expression of all the clustered genes

Figure 5: The comparison between the transcriptional state between Time *n* vs. Time 0. (Dashed) is the transcriptional State of time 0, and (Solid) is the transcriptional state at time n.

Figure 6: A comparison between the objective functions between the entire set of genes, 529 informative probes, and 529 randomly selected probes

Figure 7: Proposed gross model and the simulated drug activity. The feedback term (s + b)(s + c) associated with G(S) implies a two compartment system with an associated degradation term. K(s) was found to be constant

Figure 8: Temporal gene expression data of genes in twelve key functional categories

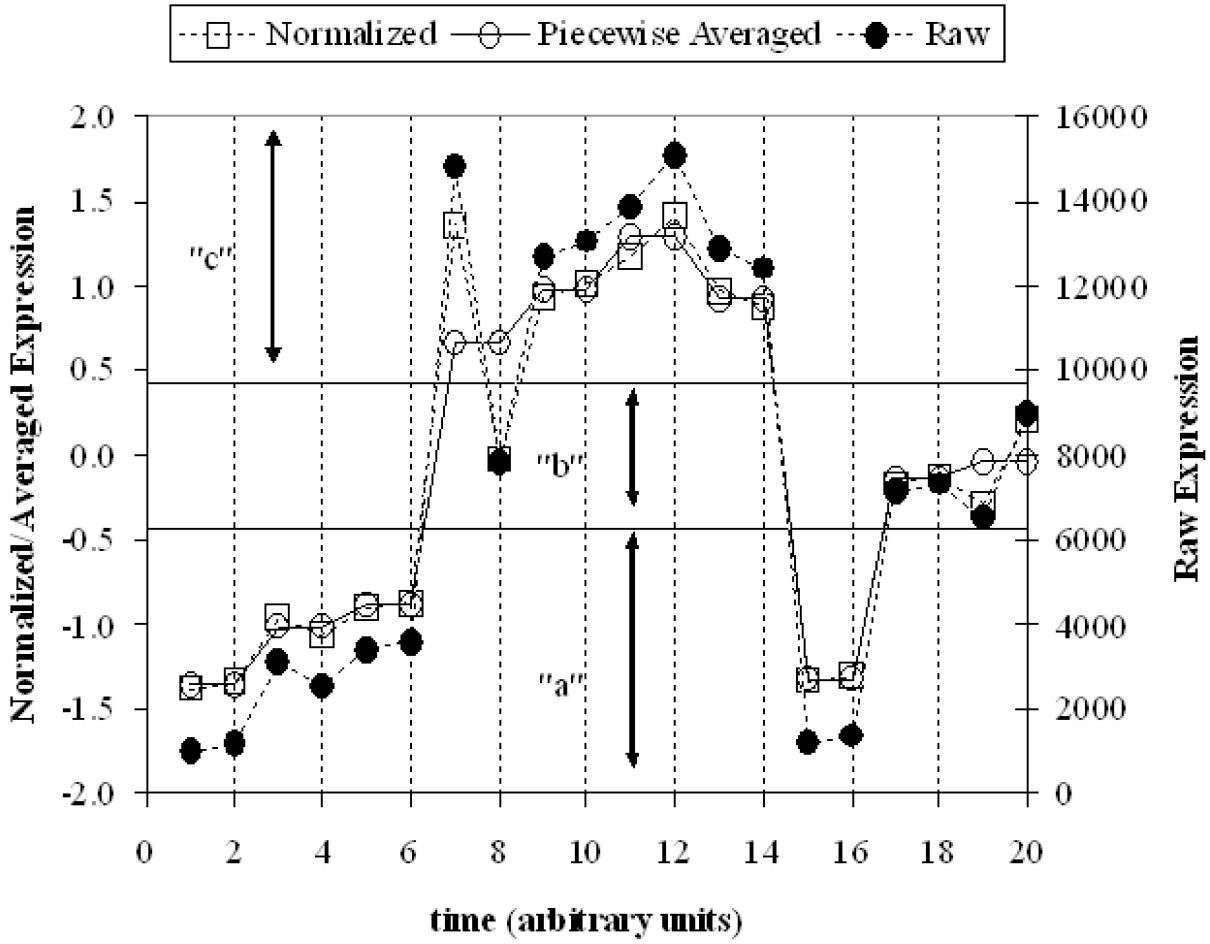


Figure 1

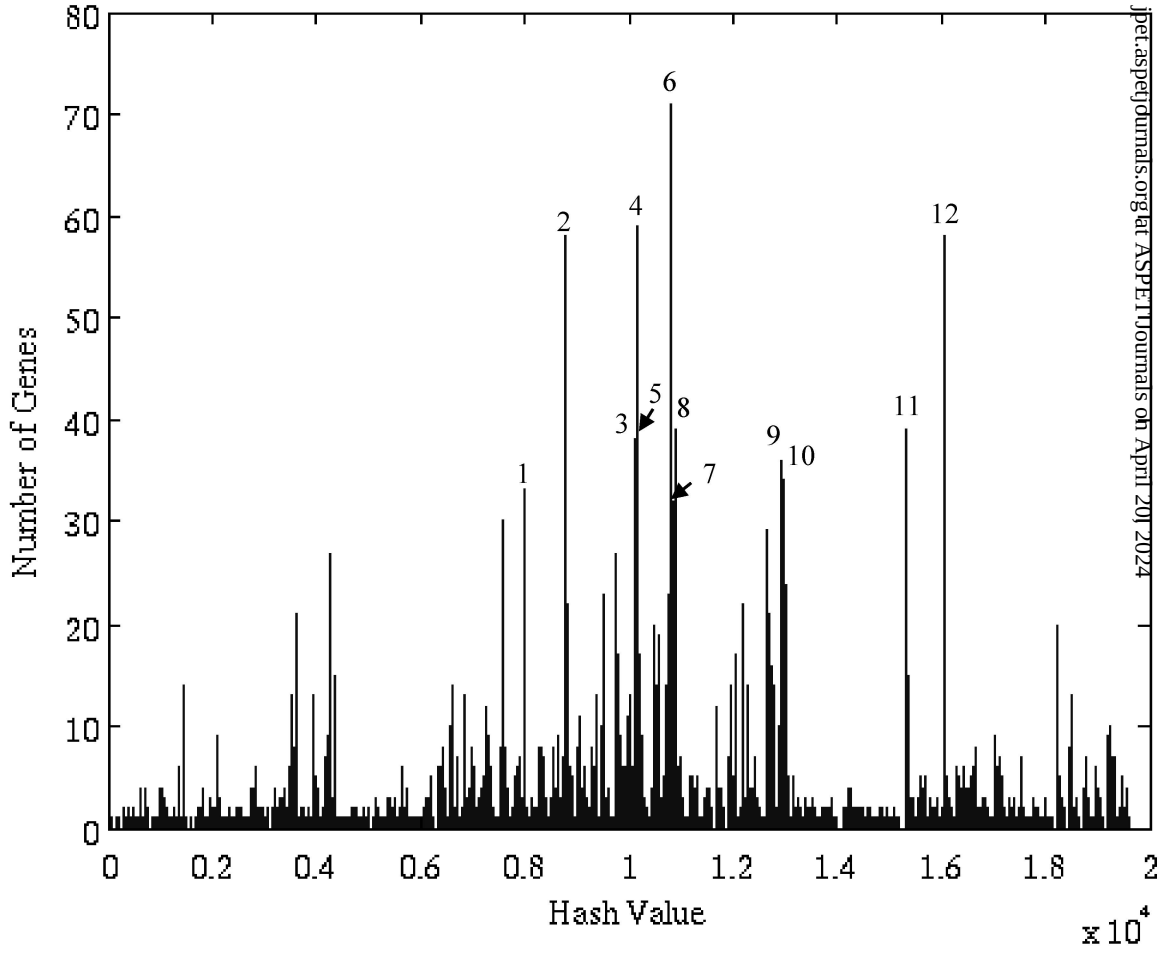


Figure 2

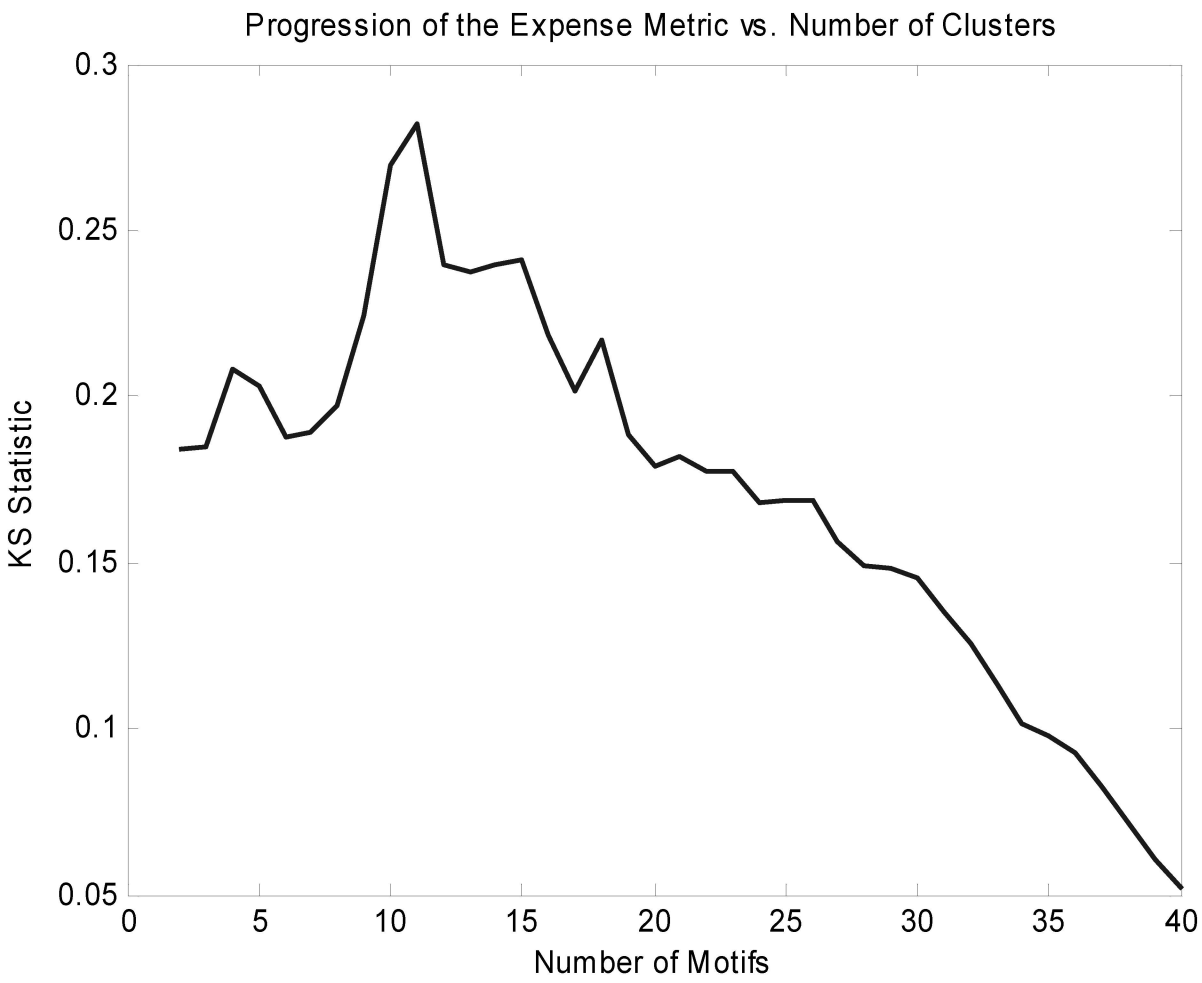


Figure 3

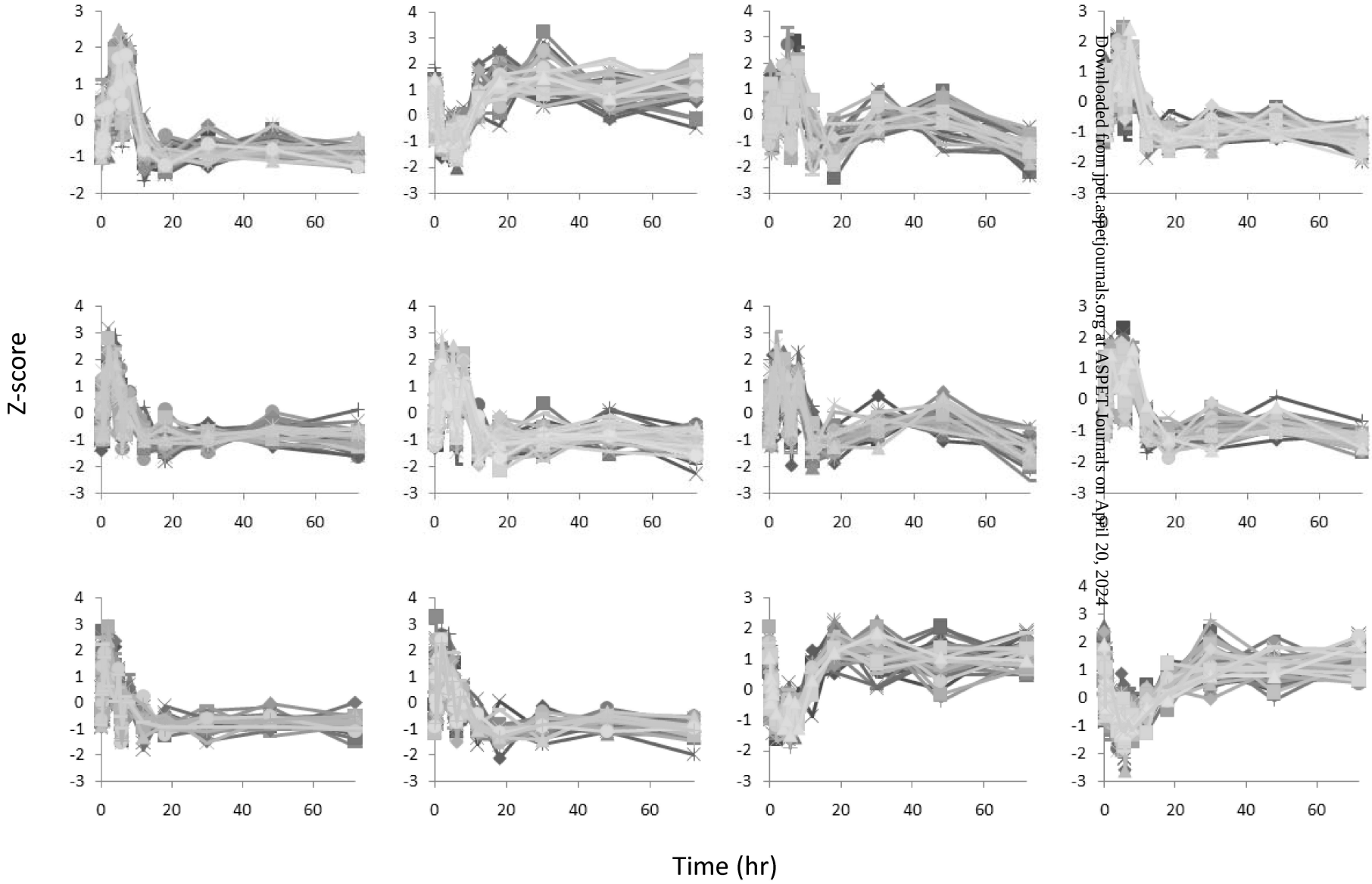


Figure 4

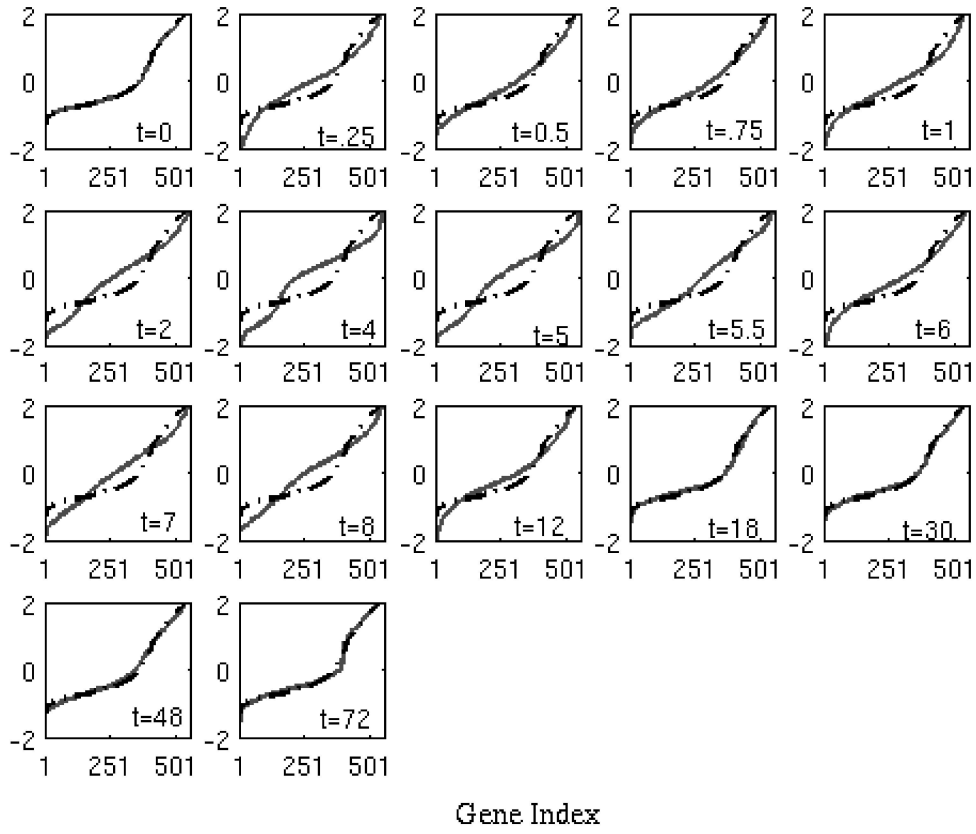


Figure 5

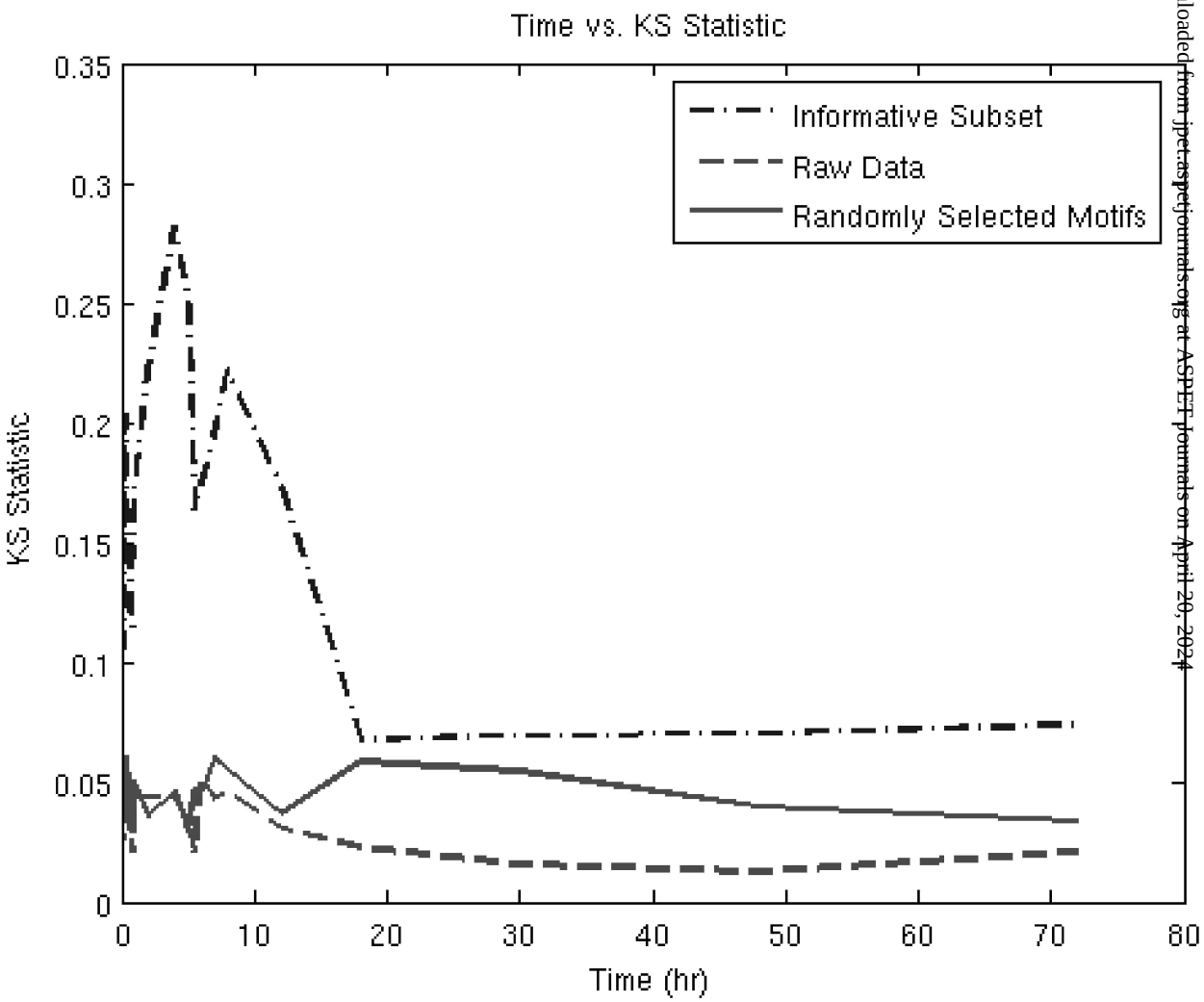
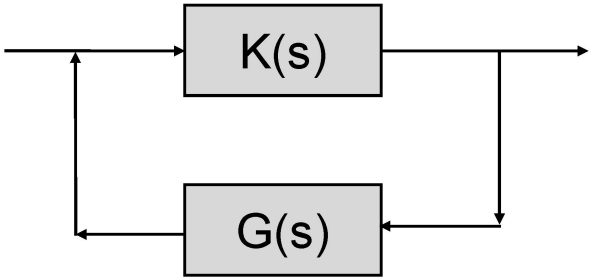


Figure 6



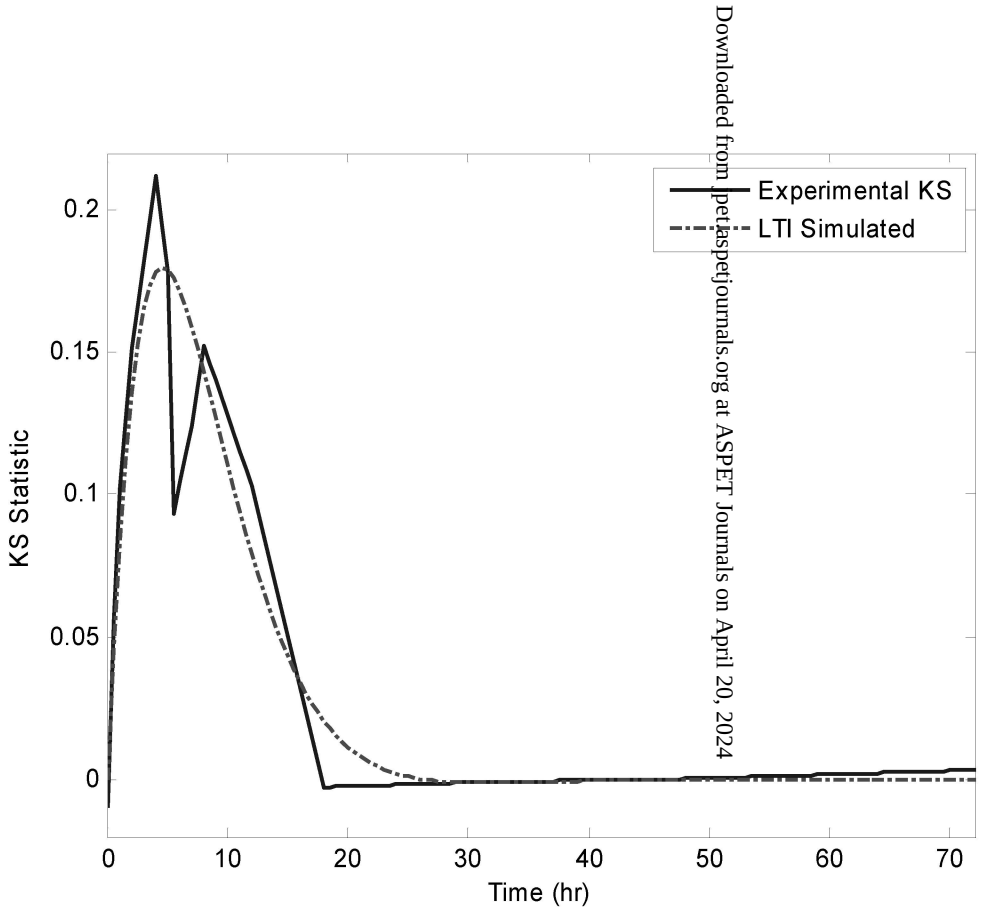


Figure 7

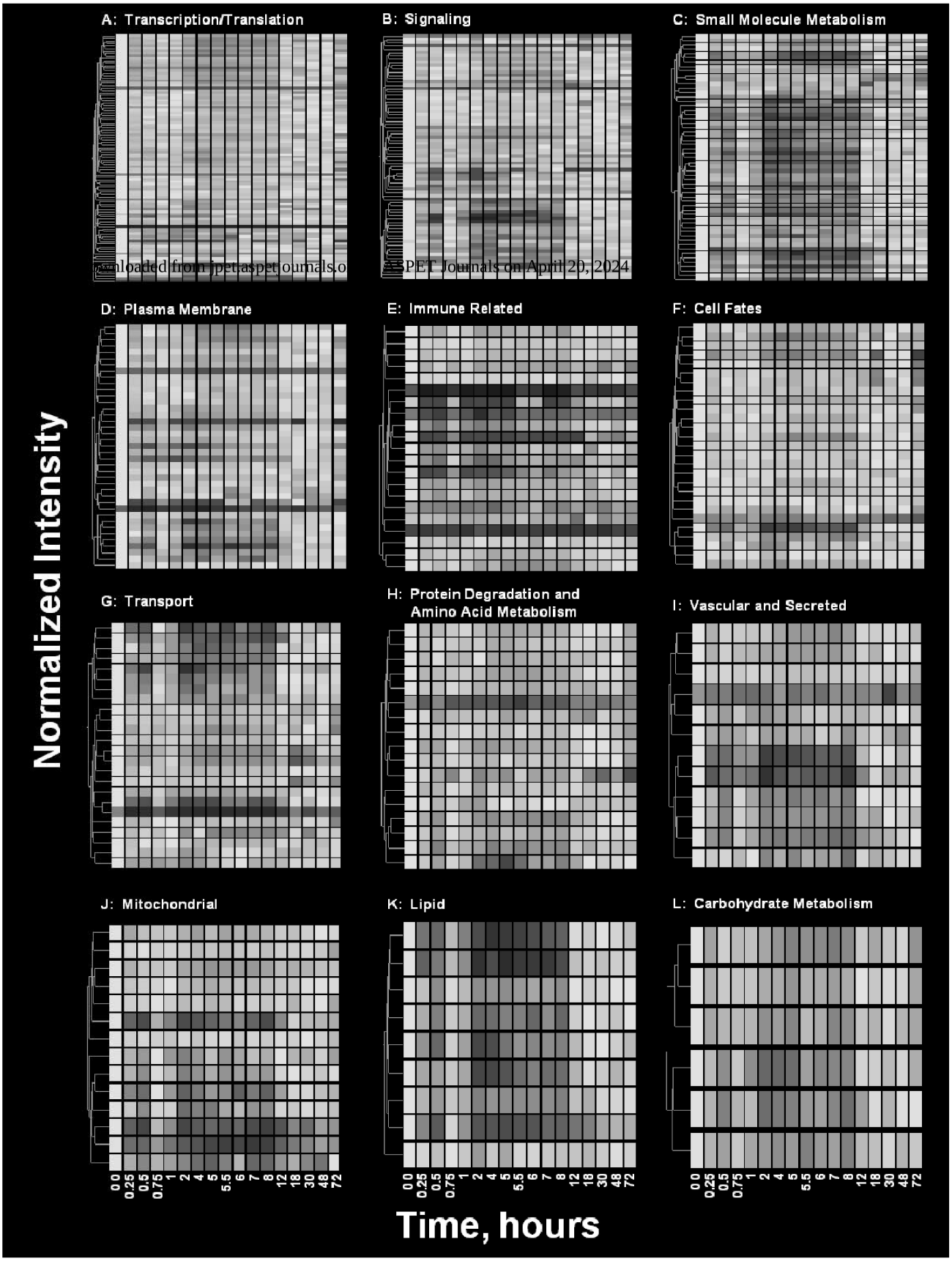


Figure 8