Timing of gene expression responses to environmental changes

Gal Chechik *, Daphne Koller,
Computer Science Department, Stanford University
Stanford, CA, 94305-9010, {*gal,koller*}*@cs.stanford.edu*

August 1, 2008

Abstract

Cells respond to environmental perturbations with changes in their gene expression that are coordinated in magnitude and time. Timing information about individual genes, rather than clusters, provides a refined way to view and analyze responses, but is hard to estimate accurately.

To analyze response timing of individual genes, we developed a parametric model that captures the typical temporal responses: an abrupt early response followed by a second transition to a steady state. This *impulse* model explicitly represents natural temporal properties such as the onset and the offset time, and can be estimated robustly, as demonstrated by its superior ability to impute missing values in gene expression data.

Using response time of individual genes, we identify relations between gene function and their response timing, showing, for example, how cytosolic ribosomal genes are only repressed after mitochondrial ribosom is activated. We further demonstrate a strong relation between the binding affinity of a transcription factor and the activation timing of its targets, suggesting that graded binding affinities could be a widely used mech anism for controlling expression timing.

*Current Address for correspondence: Google Inc, 1600 Amphitheater Parkway, Mountain View CA, 94043

Introduction

Over the past few years, significant progress has been made in mapping different components of the cellular architecture: protein complexes, functional modules, and even more complex pathways and cellular networks. However, the static set of components and their interactions tells only part of the story. In reality, cells continuously reconfigure their activity to adapt to their fluctuating environment, and activate different parts of their pathways in a dynamic way. Obtaining insight into the cellular dynamics is a significant challenge, primarily because data measuring aspects of the cell's activity over different points in time is hard to obtain, especially at a genome-wide scale.

Arguably, the main data so far that have provided a genome-wide view into the cell's dynamics are measurement of gene expression profiles taken over a time course, following a perturbation to the cell's environment. Although these measurements probe only a single level of the cellular control hierarchy, the availability of transcription data under multiple conditions could provide significant insights into dynamics of cellular control. With these data, we try to understand the role that expression timing plays in cellular responses, to map those genes and modules that are expressed in a timely manner and to identify molecular mechanisms that control timing.

Unfortunately, gene expression time courses are hard to interpret: they are notoriously noisy, often measured at irregular intervals, and these intervals differ from one experiment to the other. Thus, with the exception of cell cycle data, much of the analysis of gene expression profiles has ignored their temporal aspects, using these data primarily to identify genes that share common responses across experiments, and to associate genes with various cellular processes based on their response profiles.

Some papers do attempt to model the dynamics of expression time courses (see Androulakis et al., 2007, for a recent survey). Several approaches (Zhao et al., 2001; Alter et al., 2000; Shedden and Cooper, 2002; Wichert et al., 2004) have focused on capturing the dynamics of cell cycle time courses; these methods are tailored to the sinusoidal transcriptional patterns in the cell cycle, and do not generalize to other types of time series. In the more general setting, Bar-Joseph and other researchers (Bar-Joseph et al., 2003; Luan and Li, 2003; Simon et al., 2005; Storey et al., 2005; Ma et al., 2006) showed how splines can be used to encode continuous gene expression profiles, and successfully impute missing values and align "similar" expression profiles that exhibit different temporal properties. Some methods (Qian et al., 2001; Balasubramaniyan et al., 2005; Ernst et al., 2005) have defined "shape-based" similarity metrics for gene expression time courses, for the purpose of gene clustering, but without attempting to extract or evaluate specific timing properties. Other approaches (Holter et al., 2001; Ramoni et al., 2002; Schliep et al., 2003; Perrin et al., 2003; Zou and Conzen, 2005) use a probabilistic or regression-based time series model to capture the temporal dynamics of gene expression data. These approaches all use generic function representation, capable of capturing a broad family of response profiles, and hence tend to over-fit the data more easily. As a consequence, the

parameters of the model are typically estimated using clusters of genes, possibly obscuring finer-grained signal. Most importantly, however, these methods do not easily provide an approach for extracting biologically meaningful timing aspects of the responses in individual time courses, and compare these timing aspects across different conditions.

In this paper we propose a parametric approach that identifies interpretable timing properties of mRNA profiles, and use them to characterize the timing of cellular responses. The idea is to fit any given time course with a function that is parametrized with biologically meaningful and easily interpretable parameters.

Specifically, we describe a phenomenological model for encoding a gene's continuous transcriptional profile over time. The model is designed to capture the typical *impulse*-like response to an environmental perturbation such as changing media or stress condition: transition to a temporary level followed by a second transition to a new steady state. Thus, we define the model in terms of meaningful aspects of the response: its *onset* and *offset* times, the slope of the response, and the short term and long term response magnitudes.

We evaluate the model on a broad compendium of 481 measurements in *S. cerevisiae*, comprising 76 different gene expression time courses following diverse environmental perturbations. We demonstrate that the impulse model is rich enough to capture a wide variety of expression behaviors and at the same time robust enough to be learned from sparse data. We then show how we can use the biologically meaningful parameters that we extract from the impulse form to shed light on the cell's transcriptional response to environmental changes.

Results

An *impulse* model of responses to changes

When subjected to an abrupt change in the environmental condition, a cell typically responds by increasing the activity level of certain sets of genes and decreasing the activity level of others. In many cases, the expression level changes temporarily, exhibiting a sharp increase or decrease, and later changes again, reaching a new steady state which is often different from the original "resting" state (Fig. 1). This two-step behavior is widely observed in multiple systems, from yeast (Holter et al., 2000; Ernst et al., 2005) to human (Ramoni et al., 2002), reflecting two types of adaptive responses. First, the cell actively reconfigures some processes, typically involving both generic emergency responses and specialized processes that the cell recruits. Then, the cell achieves a new homeostasis in its new environment.

We propose an *impulse model* designed to encode such two-transition behavior, allowing us to compactly represent the relevant aspects of expression responses to environmental changes. The impulse model encodes this behavior as a product of two sigmoid functions, one that captures the onset response, and another that models the offset. Importantly, this model allows for a sustained expression level different from the resting state. The model function has six

free parameters $\theta = [h_0, h_1, h_2, t_1, t_2, \beta]$, (shown in Fig. 1(A)). Three amplitude (height) parameters determine the initial amplitude (h_0), the peak amplitude (h_1), and the steady state amplitude (h_2). The onset time t_1 is the time of first transition (where rise or fall is maximal) and the offset t_2 is the time of second transition. Finally, the slope parameter β is the slope of both first and second transitions. Formally, the model has the following parametric form:

$$\begin{aligned} f_\theta(x) &= \frac{1}{h_1} \cdot s(x, t1, h0, +\beta) \cdot s(x, t2, h2, -\beta) \qquad (1) \\ s(x, t, h, \beta) &= h + (h_1 - h) S(+\beta, t) \\ S(\beta, t) &= \frac{1}{1 + e^{-\beta(x-t)}} \end{aligned}$$

What type of profiles can the impulse model capture? It is designed for modeling temporal profiles that have at most two significant changes in expression levels. Examples of such profiles are depicted in Fig. 1(B), where the impulse model was fit to actual expression measurements of yeast genes. The impulse model is not appropriate for encoding periodic behavior with multiple peaks, such as the characteristic behavior of the cell cycle (like the well-studied data of Spellman et al. (1998)). Thus, the impulse model is best-suited for capturing a one-time response to some external signal such as an environmental disturbance.

The parameters of the model are learned by minimizing a squared error to fit measured data. Given a set of expression measurements $\{e_1, \ldots, e_n\}$ at time points $\{t_1, ..., t_n\}$, we search for the set of impulse parameters θ that minimize the squared prediction error $\min_\theta \sum_i (f_\theta(t_i) - e_i)^2$. We find the (locally) optimal parameters using a conjugate gradient ascent procedure, repeated 100 times with different starting points (see supplemental Methodsonline).

Gene expression measurements are notoriously noisy and hard to model, especially on the level of individual genes. We systematically evaluated the properties of the impulse model using a diverse set of 76 conditions. First, we found the model to be remarkably robust to both timing noise and to expression level noise. Furthermore, we estimated the model's coverage — the fraction of genes that can be well-fit with the model — showing that up to 95% of the genes are well described by the impulse model , depending on the condition. Finally, we estimated the extent to which genes had a particularly impulse-like response, showing that, on average, 35% of the genes have an impulse-like response profile. All these resutls are provided in supplemental methods online Chechik and Koller (2008).

Imputing missing values

The impulse model is a continuous function that provides an estimate for a gene's expression measurement at each point in time. We show that the impulse model can accurately predict the value of missing expression measurements. Imputing missing values is an important problem in gene expression data analysis,

hence the success of our impulse model at this task is both a validation of the model, and one of its applications.

We applied the impulse model to a compendium of 76 gene expression time courses in *S. cerevisiae*, which measure the response of yeast to different environment stress conditions and changing media (DeRisi et al., 1997; Gasch et al., 2000, 2001; Causton et al., 2001; Zakrzewska et al., 2004; Lai et al., 2005; Kitagawa et al., 2005; Mercier et al., 2005). Time courses had between 5 and 10 measurements (the full list of data sets and time courses is in Supplemental Table 1 online Chechik and Koller (2008))

We evaluated the performance of the model on the imputation task in two ways: using information only at the level of individual genes; and incorporating information from other, similar genes.

Using individual genes

First, we considered the ability of an impulse model to estimate the value of an unmeasured expression value for a gene, given the other expression measurements for that gene alone. For a given gene, we fit an impulse model to all measurements except a single held out time point, and used the resulting function to estimate the expression value at the held out measurement. We compared this value to the measured held-out value, and computed the error. We repeated this experiment for all 6209 genes in our compendium and all measurements, and computed the mean prediction error. For comparison, we applied the same procedure using other methods for function estimation, including both interpolation methods such as interpolating splines and cubic-Hermite polynomials, and fitting methods using polynomials of degrees two to five, and smoothing splines. All of these methods used information at the level of single genes only, using measurements taken at all available time point to predict the value in a single hidden time point. The prediction of the impulse model are significantly superior to all the other methods Fig. 2(A).

Fig. 2(B) shows a scatter plot of average prediction error for each of the 76 conditions, as obtained with the impulse model and the cubic-Hermite (CH, the second best predictor). It shows that the impulse model is particularly better at fitting time courses with a small number of points, suggesting that it avoids over-fitting more effectively.

Interestingly, a comparison to a third order polynomial yields similar results. This similarity suggests that even though the impulse model has 6 free parameters, it avoids over-fitting better than a model with 4 free parameters. The reason is that polynomials are generic function approximators, capable of fitting any function, hence could predict fits that are highly unlikely for gene expression timecourses. In comparison, the impulse model focuses on a restricted set of behaviors, and hence uses the domain-specific knowledge to avoid large mistakes.

This effect can be understood by comparing the actual functions learned by the different fitting procedures. Fig. 2(D)-(F) compares the fits to a particular gene expression profile for three methods: polynomials of degree 2 and 3, and

the impulse model. The descriptive power of the 2^{nd} order polynomial is too limited, leading to a "flat" curve that changes little in time. On the other hand, the 3^{rd} degree polynomial is too expressive, and over-fits for several time-points. Conversely, the impulse model, despite having a larger number of free parameters, successfully avoids over-fitting the measurements.

Using whole genome information

A valuable source of information When imputing missing values is the similarity in expression profiles between different genes. Two approaches are commonly used for taking this information into account. First, missing values are inferred from neighboring genes, where the neighborhood is based on the observed measurements. Second, genes are clustered and the cluster profiles are then used for imputing the missing values. We compare the performance of the impulse model with two standard methods that take these two approaches.

For the first evaluation, we follow the approach of Troyanskaya *et al.* (Troyanskaya et al., 2001) and use profiles of similar genes to complete missing measurements. Troyanskaya *et al.*, in their KNN-impute procedure, propose a k-nearest neighbor procedure, estimating the value of a time t measurement for gene g as the average of the time t expression values measured for the k genes most similar to g. KNN-impute uses a Euclidean distance over the vector of expression measurements to find the nearest neighbors. To evaluate the gain in using the impulse model we applied the same procedure, but using the values predicted by the impulse model fit, rather than the raw original measurements. Specifically, we hid a randomly selected single time point in the expression profile of each gene, and used the remaining measurements to estimate the left-out values (see Methodsonline).

Fig. 2(C) compares the median error obtained with the two distance measures across 76 conditions. Using the impulse model reduced the error in 64 out of 76 conditions, yielding an average error reduction of 20% of the KNN-impute. This difference was highly significant (paired t-test: $p < 3 \times 10^{-6}$). The analysis was repeated for $k = 10$ and $k = 20$, with almost identical results (not shown).

Bar-Joseph et al. (2003) used another approach for utilizing similarity of expression profiles across genes. They cluster genes and train a model based on approximating splines for cluster profiles. We compared this method with *Impulse-KNN* over the same data set described above. We used code supplied by Bar-Joseph, and selected values of the parameters that performed well in the experiments described by Bar-Joseph *et al.*. We used 10 clusters, since we found that this number of clusters captures well most of the structure in the data. The *Impulse-KNN* model outperformed spline-based clustering by 35% on average. Results are not shown due to space constraints.

Temporal patterns of response to changes

The impulse form directly provides meaningful parameters that characterize the shape of the response profile, including the response onset, offset and profile

peak. We chose to focus on the onset response time, since it directly captures the timing at which the cell initiates the production of a gene's mRNA, and this timing could be critical to the survival of the organism upon an environmental change. We therefore extracted the onset of every response profile, and used these timing data to explore the relationship between response onset and gene function.

To illustrate the insights arising from this type of analysis, we can consider the timing patterns arising when the cell is exposed to diamide (Gasch et al., 2000). Here, we can see that genes involved in *gene expression* respond at a wide range of delays (Fig. 3(A)). Looking at three main subsets of this group, we find that genes that are involved in *RNA processing* typically respond earlier than the other genes; transcription genes also respond early, and translation is last. Interestingly, translation occurs in two peaks, one observed early ($\sim$7 minutes) and a second occurring much later ($\sim$18) minutes.

To understand this phenomenon better, we look into the distribution of onset times and peak responses of all ribosomal genes under diamide exposure. A finer breakdown of the set of ribosomoal genes reveals that the vast majority of the early onset events correspond to induction of the mitochondrial ribosome, whereas the later events represent the repression of the cytosolic ribosome (see Fig. 4). We note that previous studies of these data (Gasch et al., 2000; Simon et al., 2005) have noted the differential expression of the ribosomal genes: while most cytosolic translation is repressed, the mitochondrial ribosome is induced in order to handle the oxidative stress caused by diamide. However, our onset timing analysis provides an additional dimension to this standard result, demonstrating that there is also a difference in the timing of these two events. We hypothesize that the reason for this delay is that upregulation and translation of mitochondrial genes is required to deal with the stress. Hence, cytosolic ribosomal genes can only be repressed after translation of mitochondrial genes is completed.

The data in Fig. 4 also reveals a fairly large group of cytosolic ribosomal genes that are repressed considerably earlier than the bulk of the genes in this category (see Supplemental Table 3). An in-depth investigation of these two groups of genes shows two interesting trends. First, in the early group, many of the genes (10 out of 33) are not ribosomal components but are more likely required for creation of ribosomes and for RNA processing or translational fidelity; by comparison, such genes are a small proportion of the late group (3 out of 115, $p < 10^{-6}$). One hypothesis is that the cell first represses accessory proteins, whereas the structural components are only shut off at the end, giving enough time for translation of the mitochondrial ribosome, as well as any other proteins necessary for the cell's immediate response. As a second trend, for the large ribosomal subunit, we see nine genes in the early group that code for the same component as a gene in the later group (for example, *RPL13A* shuts down early, whereas *RPL13B* shuts down later). The only case where both copies are shut down early is *RPL41A* and *RPL41B*, which code for a non-essential component of the ribosome. An interesting hypothesis is that, to conserve resources, the cell begins by shutting off one copy of each component, and only then shuts

down the other. The situation is a little less clear with the small subunit, where three components have both copies shut down early; however, these are not in the central part of the ribosome. It would be interesting to understand whether and why these components are not required during the transition phase.

To generalize this analysis and identify other functions whose RNA levels are carefully timed, we looked at the distribution of onsets across genes grouped by their GO associations. In each condition, we then searched for GO categories whose onsets are significantly different from a baseline distribution of onsets. A relevant baseline should contain genes of similar (but not identical) functions. We defined a separate baseline for each category using all genes from sibling categories in the GO hierarchy (other children of its parent category). For each GO category and each condition, we calculated a Wilcoxon score to quantify how significantly its gene onsets appear earlier or later than the baseline onsets. This comparison provides a tool for identifying sub-functions that are controlled in time. We found 151 sub-categories that exhibited highly significant (Wilcoxon test, $p < 10^{-5}$, Bonferroni corrected) onset differences at least in one condition (Chechik and Koller (2008)).

Fig. 3(B) shows another example, for the main sub categories of *intracellular organelle part*, under exposure to Acid (Causton et al., 2001). Mitochondrial genes are again regulated significantly earlier, and so are cytoskeletal genes, while a larger fraction of chromosomal respond late. Ribosomal genes again have two peaks, and these correspond again to mitochondrial and cytosolic ribosome; indeed, as we discuss below, this distinction is found across a variety of conditions. Here, vacuolar genes also appear to have two distinct peaks, with 53 genes responding before $t = 12$ minutes and 20 genes responding after. Relative to the late vacuolar genes, we find that the early vacuolar genes are enriched for *vacuolar membrane* (hypergeometric $p < 10^{-15}$).

We can also utilize our timing analysis to construct a system-level "response timeline", by looking at how multiple functional categories are ordered in time. Under each condition, we calculated the ordering score for every pair of GO categories, and used these ordering scores to identify sets of categories that are regulated in a timing-distinct manner (see Methodsonline). As one example, we consider the onset timing extracted from the responses to DNA-damaging gamma irradiation (Gasch et al., 2001). Fig. 3(B) plots the median peak and median onset time for each of the top four timed categories in the *cellular-component* hierarchy. First, genes of the nucleolus (a sub-organelle of the cell nucleus) are repressed, followed by repression of ribonucleoproteins, then cytoplasmic proteins. Finally, membrane proteins are activated. A similar analysis on annotations in the *molecular function* and *biological processes* hierarchies in the same condition (Supplementary Figures online Chechik and Koller (2008)), is consistent with this view: The biological processes of *ribosome biogenesis and assembly* (which takes place at the nucleolus) are repressed first, followed by the activation of the *localization* and *transport* genes (processes that take place at cytoplasm and membranes). Similarly, the molecular function *structural constituent of the ribosome* are repressed first, while multiple functions related to transport are activated later.

Also interesting is the observation that the stronger the repression of the genes in these timed categories, the earlier the onset of the repression. This phenomenon holds not only for the medians of the groups in Fig. 3(B) , but in fact the onset time is correlated with the peak response across all genes in these categories (Pearson correlation, $p < 10^{-10}$); this phenomenon holds only for genes in timed categories (the background correlation across all genes in this condition is p-value = 0.04). As one hypothesis, if a group of genes is highly detrimental to the cell (leading to a strong repression), it may be desirable to shut them off as soon as possible. In particular, if mRNA degradation mechanisms are used to decrease mRNA abundance in this condition (Keene, 2007), this finding may also suggest a sequential targeting of the RNA degradation machinery, ordered by the cell's current priorities.

Finally, we looked at functional differences in timing across multiple conditions. We counted the number of conditions in which each pair of categories is significantly timed (p-value < 0.001, Wilcoxon test, Bonferroni corrected). In general, nuclear and mitochondrial components respond earlier than cytosolic and ribosomal components. For instance, for the *cellular component* hierarchy, the mitochondrion, shown above to be activated early under exposure to diamide (Fig. 4), and acid (Fig. 3, responds significantly earlier (with $p < 10^{-3}$) than the cytosolic ribosome in 16 out of the 76 conditions tested (yielding an overall $p < 10^{-40}$, Binomial distribution with $p = 10^{-3}$,$N = 76$).

Graded binding affinity: A mechanism for controlling transcription timing

The above findings suggest that cells control the timing of transcription activation to shape their responses to environmental changes. What mechanisms could achieve fine timing control?

One possible mechanism is that sequential activation of genes is achieved by cooperative binding by several transcription factors (TFs), each activated in its turn. This hypothesis requires that TF's are themselves sequentially activated by some mechanism. A different (albeit not exclusive) mechanism is that a single transcription factor binds to multiple target genes, but with different binding affinities. Indeed, the recent work of Tanay (Tanay, 2006) shows that binding affinities, as measured in ChIP-chip data (Harbison et al., 2004), have functional consequences even in weak affinities that were previously considered insignificant. This work demonstrates that transcription binding is not an all-or-none phenomenon, and graded binding is achieved through graded sequence affinity. The reason and purpose for having a wide range of binding affinities is still unknown, but it was recently shown that gene expression in the phosphate response (PHO) pathway is tuned to different environmental phosphate levels using both binding-site affinities and chromatin structure (Lam et al., 2008).

If graded binding affinities are used for regulating the timing of gene expression, we expect the shape of a gene expression profile to depend on the strength of binding to its regulating TFs. Since binding operates as a stochastic equilibrium, the stronger the binding affinity of an activating TF to a binding

site, the higher the probability of the TF to remain bound to the corresponding promoter and recruit the transcriptional machinery, and hence the earlier the gene would be expressed on average.

To test this single-TF hypothesis, we measured how binding affinities are related to the onset time of transcription activation. Specifically, we combined whole genome binding affinity measurements (Harbison et al., 2004) with gene expression measurements as described above. We selected a subset of affinity and expression measurements that were taken in matching conditions. We collected a total of 48 affinity-expression experiment pairs (see Supplemental Table 2 online Chechik and Koller (2008), including amino acid starvation (34 TFs), exposure to acid (2 TFs), and to heat shock (12 TFs).

For each affinity-expression experiment pair, we restricted attention to genes that were differentially expressed (absolute peak response $> R$), and measured the Spearman correlation between their onset time and the binding affinity of the measured TF, using the p-value as the quantitative measure of affinity. Of course, not all genes are bound by a particular TF; we therefore wanted to restrict attention only to those genes where TF binding plausibly occurs. As discussed above, Tanay (Tanay, 2006) showed that measured binding affinity p-values are correlated with binding prediction based on sequence models, even for very weak binding, suggesting that measured weak binding may reflect actual binding rather than noise. We therefore considered the whole range of possible p-value thresholds for treating a binding event as valid (where the chance level is p-value $= 0.5$). Specifically, for a range of different affinity thresholds C, we computed the Spearman correlation between onset time and binding affinity, restricting the analysis to all genes that are both differentially expressed (crossing a threshold R) and have a binding affinity stronger than a cutoff value C. Fig. 5(A) shows the number of pairs that obtained significant Spearman correlation as a function of the affinity cutoff value C; here, we used a gene expression response threshold $R = 0.7$, chosen to maximize the number of significant pairs. The number of significant pairs peaks near p-value $= 0.50$, where 38 of 48 TF-condition pairs have a significant correlation (FDR q-value$\leq 10^{-3}$) (the optimum is actually obtained at 0.52, which is larger than the chance level 0.5, but this is likley to be due to noise, . Typically, the correlations became even stronger when limiting the analysis to more strongly expressed genes (larger values of R), but the p-values decrease due to the smaller sample size.

Fig. 5(B) -(C) visualizes the relation between binding affinity and expression onset; here, to more clearly illustrate the pattern, we used an expression cutoff of $R = 1$. We aggregated the genes in our set into four groups according to their binding affinities, and calculated the mean onset time of each group. The left panel shows the results of this analysis for the targets of MET32, a transcription factor involved in methionine biosynthesis; here, the binding affinities were measured under amino acid starvation, and the transcription onset extracted from a time course following adenine starvation (Gasch et al., 2000). A clear trend can be observed in the mean onset time as a function of MET32 binding affinity, across the whole range of relevant affinity strengths. This effect is highly significant (Spearman correlation $r = 0.14$ across 943 samples, $p < 1.9 \times 10^{-7}$,

Bonferroni corrected for 48 hypotheses). Other such trends were observed under amino acid starvation, including MET31 (Fig. 5(C)) (Spearman $r = 0.14$, Bonferroni $p < 2.5 \times 10^{-8}$), CBF1 ($p < 9.7 \times 10^{-8}$) and SFP1 ($p < 6 \times 10^{-9}$).

We also found pairs that exhibited significant negative correlations (for instance YAP1 and HSF1 under a heat shock), where higher binding affinity was associated with delayed onset time. The mechanism for such associations is unclear at this point, and could be related to competition between TFs.

This finding has two implications. First, it shows that graded binding affinities are very commonly correlated with expression timing, and could be a commonly used mechanism for controlling the timing of response onsets. Second, it suggests that even (very) weak binding affinities have a functional effect on the concerted profile of cellular expression responses.

Discussion

Environmental changes may threaten the survival of cells and force them to respond quickly and reconfigure their gene expression profiles. To respond efficiently to changing conditions, cells have to control not only the magnitude of their responses, but also their timing. Indeed, it was shown that expression timing in *E. Coli* is tightly controlled, even to the level where sequences of individual proteins are expressed in an ordered manner (Zaslaver et al., 2004; Kalir et al., 2001). It is unknown, however, if such controlled timing is to be found across multiple biological processes, and if responses are similarly timed in Eukaryotes, which have more complex hierarchy of pre- and post-transcriptional control mechanisms. Our work suggests that fine-grained control of transcriptional timing exists also in Eukaryotes.

The time course of gene expression responses often follows a typical *impulse* curve: starting with an initial abrupt response that saturates and is then followed by a relaxation to a new steady state. In this paper, we used this common behavior to build a parametric model that can be robustly fit to a single expression profile, while capturing the essential timing aspects of the response: its onset time, peak response and offset time.

Since the impulse model is tuned to typical cellular responses, it provides robust estimates of response characteristics, even when given very few samples per time course. We found that it provides superior prediction for imputing missing or corrupted measurements, both using single gene and using whole genome information. We believe that this model has other valuable uses, such as the alignment and comparison of time courses taken at different time points.

Perhaps most important, the impulse model allows us to study response timings directly. Using the distribution of onsets across functional categories, we found multiple functions that are timed differently from closely related functions. We also observed a global response pattern, roughly moving outwards from the nucleus towards the cytoplasm and membranes. Finally, we found strong correlations between the onset of responses and the binding affinity to specific transcription factors. This last finding suggests a hypothesis in which

gradual binding affinities are widely used by cells to tune the timing of expression responses, extending on recent findings in the context of specific pathways (Lam et al., 2008).

Transcriptional regulation is one mechanism in a series of hierarchical controls including regulation of mRNA, translation, and protein activation. Importantly, we note that our finding relates to overall mRNA levels, which encompass effects from both transcriptional changes and mRNA degradation. Our analysis is done purely on the timing in the change in mRNA levels, and we make no attempt to identify the cause(s) for the change. Indeed, it seems quite likely that many of the timing changes we observe result from a combination of these two regulatory mechanisms. Regardless of the mechanism, our findings suggest that the timing in fluctuations of gene expression levels is regulated in a way that optimizes for the role of the resulting protein product. For instance, the distribution of timing in Fig. 4 suggests a bifurcated response in the cytosolic ribosome: those components that are not required for translation of other protein products are repressed early, whereas the necessary components are repressed later, after fulfilling their role. Therefore, even though several regulatory phases separate mRNA levels from active protein levels, our findings support a model in which response onsets of mRNA are tuned with respect to the corresponding protein function.

The impulse model captures one kind of typical response profiles, but other typical behavior may exist, such as the periodic behavior observed due to cell cycle. Such typical behaviors can be identified by unsupervised clustering of time courses, as in (Ernst et al., 2005). As a subsequent step, one can then construct a specialized model as in this paper that utilizes biologically relevant parameters that characterize that type of response, allowing these parameters to be extracted and used in further analysis.

Impulse-shaped responses are not limited to mRNA responses to stress. Similar patterns are observed in gene expression profiles along early development (Wen et al., 1998) or in protein profiles. The modeling and visualizations techniques discussed in this paper could be usefully applied in these cases as well. Analysis of gene expression data can be used to analyze the dynamics of cellular networks, seeing how they adapt in response to changes in the cell condition. Recent work uses these data to obtain important insights into the dynamics of complex formation during the cell cycle (de Lichtenberg et al., 2005). We hope that the fine-grained timing information provided by our work will allow us to understand the reconfiguration of cellular complexes and pathways in response to environmental perturbations.

References

Alter, O., Brown, P., and Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*, **97**(18):10101–10106.

Androulakis, I., Yang, E., and Almon, R., 2007. Analysis of time-series gene expres-

sion data: Methods, challenges, and opportunities. *Annual Review of Biomedical Engineering*, **9**:205–228.

Balasubramaniyan, R., Hullermeier, E., Weskamp, N., and Kamper, J., 2005. Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics*, **21**(7):1069–77.

Bar-Joseph, Z., Gerber, G., Gifford, D., Jaakkola, T., and Simon, I., 2003. Continuous representations of time series gene expression data. *J Comput Biol*, **10**(3-4):241–256.

Causton, H., Ren, B., Koh, S., Harbison, C., Kanin, E., Jennings, E., Lee, T., True, H., Lander, E., and Young, R., *et al.*, 2001. Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell*, **12**(2):323–337.

Chechik, G. and Koller, D., 2008. Supplemental data online, http://ai.stanford.edu/ gal/research/impulse.

de Lichtenberg, U., Jensen, L., Brunak, S., and Bork, P., 2005. Dynamic Complex Formation During the Yeast Cell Cycle.

DeRisi, J., Iyer, V., and Brown, P., 1997. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science*, **278**(5338):680.

Ernst, J., Nau, G., and Bar-Joseph, Z., 2005. Clustering short time series gene expression data. *Bioinformatics*, **21**(Suppl 1):i159–i168.

Gasch, A., Huang, M., Metzner, S., Elledge, S., Botstein, D., and Brown, P., 2001. Genomic expression responses to dna-damaging agents and the regulatory role of the yeast atr homolog mec1p. *Mol Biol Cell*, **12**(10):2987–3003.

Gasch, A., Spellman, P., Kao, C., Carmel-Harel, O., Eisen, M., Storz, G., Botstein, D., and Brown, P., 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, **11**:4241–4257.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J.-B., Reynolds, D. B., Yoo, J., *et al.*, 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**:99 – 104.

Holter, N., Maritan, A., Cieplak, M., Fedoroff, N., and Banavar, J. R., 2001. Dynamic modelling of gene expression data. *Proc Natl Acad Sci U S A*, **98**(4):1693–1698.

Holter, N., Mitra, M., Maritan, A., M. Cieplak, J. B., and Fedoroff, N., 2000. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proc Natl Acad Sci U S A*, **97**(15):8409–8414.

Kalir, S., McClure, J., Pabbaraju, K., Southward, C., Ronen, M., Leibler, S., Surette, M., and Alon, U., 2001. Ordering Genes in a Flagella Pathway by Analysis of Expression Kinetics from Living Bacteria.

Keene, J., 2007. RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet*, **8**:533–543.

Kitagawa, E., Akama, K., and Iwahashi, H., 2005. Effects of iodine on global gene expression in saccharomyces cerevisiae. *Biosci Biotechnol Biochem*, **69**(12):2285–2293.

Lai, L., Kosorukoff, A., Burke, P., and Kwast, K., 2005. Dynamical remodeling of the transcriptome during short-term anaerobiosis in saccharomyces cerevisiae: differential response and role of msn2 and/or msn4 and other factors in galactose and glucose media. *Mol Cell Biol*, **25**(10):4075–91.

Lam, F., Steger, D., and O'Shea, E., 2008. Chromatin decouples promoter threshold from dynamic range. *Nature*, **453**(7192):246.

Luan, Y. and Li, H., 2003. Clustering of time-course gene expression data using a mixed-effects model with B-splines.

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J., 2006. A data-driven clustering method for time course gene expression data. *Nucleic Acids Res*, **34**(4):1261.

Mercier, G., Berthault, N., Touleimat, N., Kepes, F., Fourel, G., Gilson, E., and Dutreix, M., 2005. A haploid-specific transcriptional response to irradiation in Saccharomyces cerevisiae. *Nucleic Acids Res*, **33**(20):6635.

Perrin, B., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., and d'Alche Buc, F., 2003. Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, **19**(Suppl 2):138–148.

Qian, J., Dolled-Filhart, M., Lin, Y., Yu, H., and Gerstein, M., 2001. Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J Mol Biol*, **314**(5):1053–66.

Ramoni, M., Sebastiani, P., and Kohane, I., 2002. Cluster analysis of gene expression dynamics. *Proc Natl Acad Sci U S A*, **99**:9121–9126.

Schliep, A., Schonhuth, A., and Steinhoff, C., 2003. Using hidden Markov models to analyze gene expression time course data. *Bioinformatics*, **19**:1264–72.

Shedden, K. and Cooper, S., 2002. Analysis of cell-cycle gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proc Natl Acad Sci U S A*, **99**:4379–4384.

Simon, I., Siegfried, Z., Ernst, J., and Bar-Joseph, Z., 2005. Combined static and dynamic analysis for determining the quality of time-series expression profiles. *Nature Biotechnology*, **23**:1503–1508.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B., 1998. Comprehensive identification of cell cycle regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**:3273–3297.

Storey, J., Xiao, W., Leek, J., Tompkins, R., and Davis, R., 2005. Significance analysis of time course microarray experiments. *Proc Natl Acad Sci U S A*, **102**(36):12837.

Tanay, A., 2006. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Research*, **16**(8):962.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R., 2001. Missing value estimation methods for dna microarrays. *Bioinformatics*, **17**:520–525.

Wen, X., Fuhrman, S., Michaels, G., Carr, D., Smith, S., Barker, J., and Somogyi, R., 1998. Large-scale temporal gene expression mapping of central nervous system development. *Proc Natl Acad Sci U S A*, **95**(1):334.

Wichert, S., Fokianos, K., and Strimmer, K., 2004. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, **20**:5–20.

Zakrzewska, A., Boorsma, A., Brul, S., Hellingwerf, K., and Klis, F., 2004. Transcriptional Response of Saccharomyces cerevisiae to the Plasma Membrane-Perturbing Compound Chitosan. *Eukaryotic Cell*, **4**(4):703–715.

Zaslaver, A., Mayo, A., Rosenberg, R., Bashkin, P., Sberro, H., Tsalyuk, M., Surette, M., and Alon, U., 2004. Just-in-time transcription program in metabolic pathways. *Nat Genet*, **36**:486–491.

Zhao, L., Prentice, R., and Breeden, L., 2001. Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc Natl Acad Sci U S A*, **98**:5631–5636.

Zou, M. and Conzen, S., 2005. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, **21**(1):71–79.

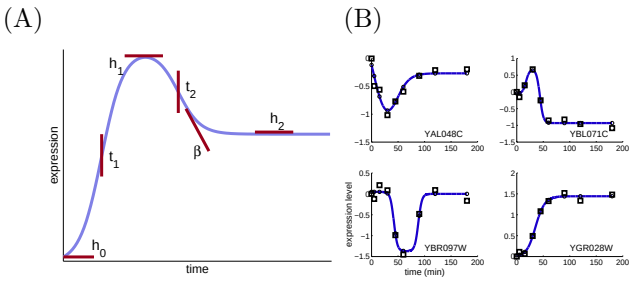

Figure 1: **The impulse model**. **(A)** The six parameters of the impulse model. **(B)** Examples of impulse model fit (solid line) to gene expression (squares) in response to $1M$ sorbitol.

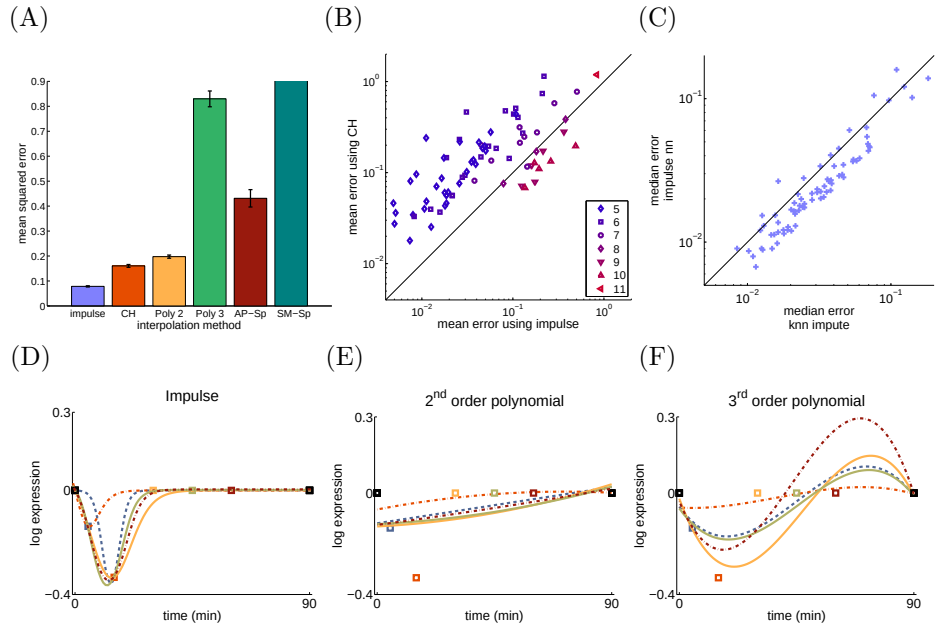

Figure 2: **Imputing missing values. (A) Mean squared error over single-gene-imputation**. Using the leave-one-out procedure described in the text, and comparing: impulse model, cubic Hermite (CH), 2^{nd} and 3^{rd} order polynomials, approximating splines and smoothing splines. Error is the average over 6209 genes in 76 conditions. Error bars denote the standard error of the mean across the 76 conditions. **(B) Scatter plot of the mean error for the impulse model and cubic-Hermite from (A)**. Each point corresponds to a different condition, and its shape shows the number of time point measurements in that condition. The impulse model provides superior fits, especially in conditions with a small number of time points. Note that the figure is in log-log scale, demonstrating that the impulse model is superior across the full range of errors. **(C) Whole genome imputation.** Comparison with Euclidean nearest neighbor KNN-impute . **(D)-(F) Comparison of leave-one-out fits to a gene expression profile**. Squares denote measurements, which are the same for all three panels. For each method, 5 curves are shown, each corresponding to a fit performed with a different single measurement that was left out during the fit. The color of each curve corresponds to the color of the hidden value (square marker).

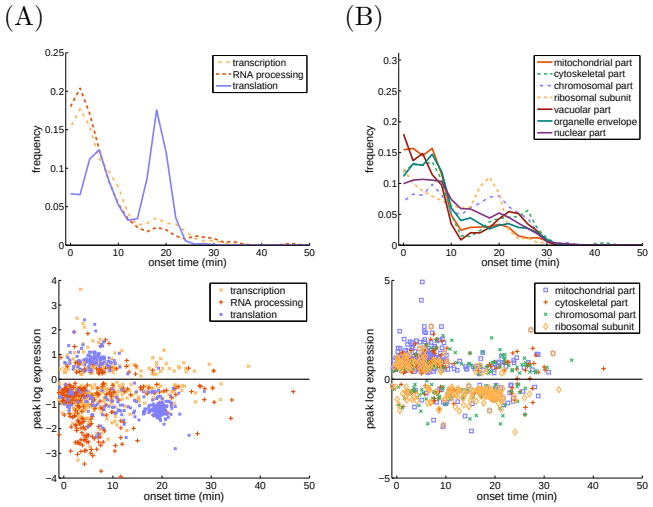

Figure 3: **Distribution of onset time and peak responses in sibling GO categories**. **(A)** Subclasses of the *gene expression* GO category, under exposure to diamide. **(B)** Subclasses of *intracellular organelle part*; Exposure to acid. Only 4 subclasses shown to reduce clutter.

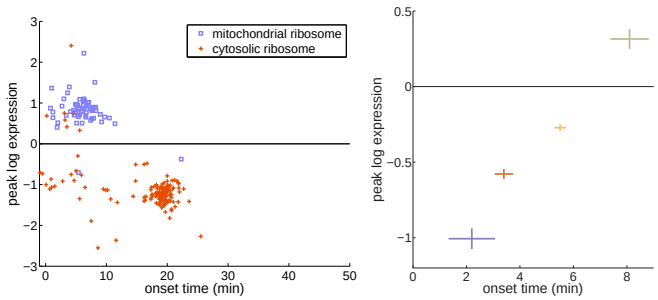

Figure 4: **(A) Onset time and peak response of Ribosomal gene responses to diamide**. **(B) The timeline of functional responses following gamma irradiation.** Crosses denote the median peak and onset of all genes in the corresponding GO category. Bar lengths denote the standard error of the mean per group (see Methods).

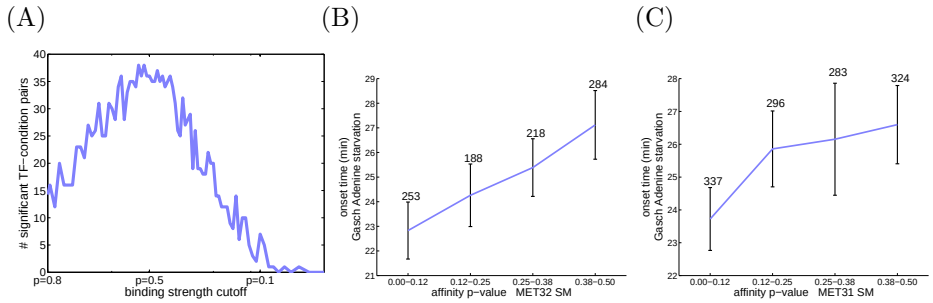

Figure 5: **(A) Number of significant TF+condition pairs as a function of binding strength considered**. The peak is achieved at p-value $= 0.52$, (38 out of 48). A p-value of 0.5 corresponds to by-chance binding. **(B)-(C) Mean onset time across genes grouped by their binding affinity. (B)** Binding of MET31 measured under amino acid starvation and expression measured under adenine starvation, Spearman $r = 0.14$ $p < . \times 10^{-9}$ **(C)** MET32, same conditions. Spearman $r = 0.13$, $p < . \times 10^{-8}$ Error bars denote standard deviations, numbers denote group sizes.