Temporal Dynamics of Regulatory Networks in *Drosophila melanogaster* Embryogenesis

Rogerio Candeias[1,2,3], Manolis Kellis[2,3]

[1]PhD Program in Computational Biology of Instituto Gulbenkian de Ciência, Oeiras, Portugal; [2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; [3]Broad Institute, Cambridge, Massachusetts 02141, USA.

Abstract

Dynamic modeling of regulatory networks that control gene expression requires temporal information on the activation/repression latencies of regulator and target pairs, which have been experimentally inaccessible at the genome scale. We developed a new discretization method using multi-step functions, to systematically infer latency information for individual edges of a large-scale regulatory network from whole-genome time-course expression profiles. Our method has wider applicability and shows increased accuracy relative to previous approaches such as Pearson or Spearman correlation. It also exhibits good predictive power of expression co-localization for regulator/target pairs benchmarked against the ImaGO annotation for *Drosophila melanogaster*.

Application of this method to *D. melanogaster* led to several new insights on its network dynamics. First, the measured delays are significantly longer than expected by chance, and are specific to *D. melanogaster*. They are not found in yeast, suggesting that they are likely relevant to animal genomes and developmental processes. Second, we found that regulator binding site multiplicity on the target promoter region is related to an increased latency, which is consistent with a slower activation associated with protein accumulation. Third, regulators of the same functional category were more likely to show similar delay distributions, suggesting different time-scales may be at play for different biological processes. Fourth, the combinatorial sum of multiple regulators is able to better explain the target expression profile than expected. Lastly, network motifs such as transcription cascades and feed-forward loops showed characteristic time delay distributions, suggesting both connectivity and dynamics contribute to the function of networks motifs.

Overall, temporal information has the potential to fundamentally change the way we think about gene regulatory networks, and the dynamic network of 21,231 temporally-decorated edges

provided here enables the study of information flow and developmental dynamics at the systems-level.

Author Summary

Embryonic development is an extremely time-dependent process requiring a high level of synchronization and accuracy. A complex mixture of several mechanisms, including a network of regulators that interact with and control the expression of target genes, is responsible for this. Several models exist for small subsets of these processes based on precise experimental measurements but no large-scale analysis has been possible given the experimental intractability of directly measuring dynamics across a complete network. Using a new discretization procedure that fits multi-step functions to genome-wide time-course expression data during embryogenesis, we were able to discover temporal delays of the edges of the *Drosophila* regulatory network. Less complex species that are typically used for dynamic modeling did not display such latencies. We found that temporal delays were longer for promoters with multiple weak DNA binding sites, consistent with a mechanistic model of transcription factor protein accumulation in promoter regions of target genes that we propose, providing a possible mechanism to fine-tune the time until expression of a given gene.

Using the inferred edge delays, we measured the time until expression of the most downstream target gene for two common network motifs, the transcription cascade and the feed-forward loop. We found that different activation/repression subtypes of these motifs show different delay distributions, suggesting that dynamics likely play an important part in determining network motif function that has not been systematically studied before.

Introduction

Gene expression induction by a transcription factor is a complex stochastic interplay of multiple players at different time-scales. Variable delays might exist between the induction of a transcription factor and the eventual expression change of its target genes, arising from a number of mechanisms, including: different affinity of regulatory motifs to the promoter region of the target gene (1), cooperative and antagonistic effects between multiple transcription factors (2), chromatin accessibility of promoter regions (3) and other effects. While several approaches for network inference have been developed using expression data from time-course microarray experiments (4-6), chromatin IP data for several transcription factors (7) or computational prediction of regulator targets based on regulatory motif instances (8), these networks have

typically remained static and devoid of any temporal information on the latency of individual edges.

In contrast, dynamic models for network motifs (9) have been extensively developed to explain both their properties and their function, similar to the dynamics of electronic circuits. For example, a negative feedback loop can lead to oscillatory or damping behavior, depending on the relative latencies of forward-pointing and backward-pointing edges. Similarly, a feed-forward loop can lead to amplification or damping depending on the relative delays of its edges (9). These properties have been largely unexplored in regulatory networks, due to the lack of edge-level temporal delay information at the full network scale, and the lack of genome-wide network connectivity information in animal genomes. In addition, most dynamic models rely on the basic assumption that a time-scale separation exists between the dynamics the of transcription factor binding to a promoter region of a target and are dramatically faster than subsequent transcription and translation, and thus edge-specific delays in a regulatory network would not be observable. Furthermore, some methods of predicting network interactions from gene expression profiles do not incorporate temporal information directly, and instead assign causality between a regulator and its target only if they are co-expressed at a given developmental stage, while temporal latencies admit lack of co-expression despite causality.

In this paper, we seek to systematically study the properties and function of temporal delays in regulatory networks. We take advantage of the combination of datasets available for model organism *Drosophila melanogaster*, consisting of both high-quality genome-wide inference of regulatory connections (8), and high-density temporal gene expression information during embryo development (10), for a system where temporal dynamics play a key role. We present a new method for inferring delays on regulatory edges based on a two-state discretization of gene expression changes for both regulators and targets, and minimum delay causality inference. We use this method to infer delays for 31,500 regulatory edges in *Drosophila melanogaster* using 28 time-points between 1,5 and 24 hours during embryogenesis (11), a temporal resolution which allows identification of sharp transitions between high and low transcription levels. Notably, the edge delays were not found in yeast time-course expression datasets using the same methods, suggesting that edge delays have been previously overlooked because they are not as striking in organisms more typically amenable to systems biology studies.

Overall, we believe this work provides a new dynamical view of transcription regulatory networks in animal genomes that will enable more precise modeling of the embryo development and other dynamic processes in the cell and perhaps in disease progression.

Results

Discretization with Multi-Step Functions

We first sought a robust method for estimating the temporal delay between a regulator and its downstream targets during development. Previous methods have used time-lagged correlations like the cross-correlation function (CCF) which assumes expression profiles have near identical curves and variance (12), or fitting single-step functions assuming a single state-transition for every regulator and target protein (13). However, neither assumption holds for the developmental regulatory network of *Drosophila*. On one hand, many regulators are only active during a finite stage during development and thus their expression profiles will not fit a single step function. On the other hand, different regulators may be responsible for multiple state transitions of the target gene at different times, and thus the expression profiles of a target and its regulators will not necessarily correlate over their entire length.

To account for these particularities of developmental regulatory networks, we developed a new method for inferring temporal edge latencies of a regulatory network based on time-course expression profiles of regulators and target genes. Our method makes two key assumptions: (a) different genes have different expression levels for active (ON, high expression level) or inactive (OFF, low expression level) states, and (b) that multiple transitions between these states are possible for a target gene, in response to discrete transition events in the expression levels of its multiple regulators. In addition, we assume that state transitions are sharp, that intermediate levels of expression are due to stochastic noise but not biologically relevant. Given these assumptions, we developed a discretization procedure that maps the continuous expression profile of each gene (for targets and regulators alike) into a pulse function whose parameters are the average values of high and low expression states and the indices of transition times between high and low states. As the number of possible transition-sequences is finite (though exponential in the number of time-points), and the best-fit values for high and low can be uniquely determined once a transition path is given, it is possible to exhaustively enumerate all parameter assignments, evaluate their fit, and select the scoring pulse function that minimizes the root square error (see Methods). To avoid overfitting however, we apply a regularization term that limits the number of transitions by imposing an Akaike Information Criterion (AIC).

All of the unique 11.990 genes mapping to a valid Flybase identifier in the time-course microarray dataset GSE6186 available in the Gene Expression Ommibus (11), were discretized into ON and OFF states using our multi-step functions method after normalization, allowing for up to 7 state changes (S_{max} parameter) in each expression profile. We found 257 expression

profiles with no state changes (2.14%), 6655 with exactly one (55.5%), 3202 with two (26.7%), 1649 with three (13.8%) and finally, 227 with four state changes (1.9%). There were no expression profiles with more than four state changes, attesting that the chosen S_{max} parameter is high enough not to constrain model selection (Figure S1). Interestingly, expression profiles with exactly two state transitions were biased to start in a low expression state (low: 27.1%, high: 72.9%), while profiles with one and three state changes exhibited no bias (50.4% / 49.6% and 50.9% / 49.1% respectively), consistent with a bias for a single finite time of activity for two-transition profiles. The 257 expression profiles where the best-fitted model holds no state changes should correspond to genes that have either basal expression or may not be required to become active during embryogenesis.

To validate whether our assumption of discrete state transitions holds, we explicitly searched for genes that would violate it. For all the genes in the regulatory network, only 44 (0.16%) showed linearly increasing or decreasing profiles (linear regression coefficient of determination higher that 0.9), and for 43% of those the procedure finds an expression state change in the middle of the expression profile. All of these genes were target nodes in the network, not corresponding regulators, and thus should not hinder our subsequent analysis.

The expression profiles of *engrailed*, *Kruppel* and *couch potato* are presented as an example of the discretization procedure (Figure 1). In each case, the ON/OFF state assignment seems to display robustness to outlier points in the expression profile that may come from experimental error. The multi-step functions procedure predicts all these profiles to have exactly two state changes, corresponding to a single fixed interval where they are in the ON state. The sequence-derived regulatory network predicts a feed-forward loop regulatory motif with both transcription factors *engrailed* and *Kruppel* regulating *couch potato*. While the delay estimation procedure for a particular edge is independent of other connections of those two nodes, there is a coherent timing between the *engrailed* and *couch potato* edges and the sum of the delays of the other two edges.

Specific Network Delay Distribution Shift in Embryogenesis

As embryonic development is highly dependent on the correct timing of events, we postulate that the distribution of edge delays found in the true network should be different from the one found in a random network. We compared the delays found in the regulatory network predicted from motif occurrences, the experimentally-derived network of ChIP-chip protein-DNA interactions (8) and a randomized network with similar degree distribution, and assessed if the distribution of delays were significantly different (Figure 2).

The distribution of delays differs significantly between the true network and the randomized networks. The experimentally-determined regulatory network and the computationally predicted regulatory network exhibited similar distributions, which given their independence confirms the high quality of both the predicted network and the delay estimation method. The two real network distributions displayed an enrichment of delays from 2 to 6 hours and a depletion of edges having 0 to 2 and 8 to 10 hours compared to the randomized network, consistent with biologically relevant activation delays during development.

Multi-step functions were able to measure a delay for 21231 out of 27682 interactions for which there is an expression profile for both genes in the microarray. This corresponds to 76.70% of edges in the network, which is estimated to have 60% accuracy, and the measured expression profiles, implying that this method could further be useful to refine network predictions. The experimentally determined subset exhibits a similar recall of 76.55%, testament to the similar quality of computationally-predicted edges and experimentally-determined edges.

Applying the same methods to the yeast cell cycle data showed no difference between the experimentally determined regulatory network (14) and a randomized network, suggesting that temporal delays may be a feature specific to the developmental process found in animal genomes. We believe this hypothesis is worth testing in the networks of additional metazoan species as they become available. It also suggests that these delays may have been a previously overlooked feature of regulatory networks, as these are not as prevalent in species traditionally amenable to systems-level analysis.

Comparison to Other Methods

We compared the multi-step functions method to other procedures that could be used to estimate a delay between expression profiles, including CCF and sliding Pearson and Spearman correlations (Figure 3). These alternative methods always attribute a causal delay for an edge, while multi-step functions have the ability to either classify an edge as having or not having a causal effect. Sliding Pearson and Spearman correlations also suffer from a small count over-estimation of delays as the number of time-points decreases. Using the correlation coefficients and the model score of the fitted multi-step functions, we performed a Kolmogorov-Smirnov test for the distribution of true edges in the network against random pairs of expression profiles. It should be noted that for each edge in the multi-step functions both genes model scores (M_{score}) were averaged. Multi-step functions displayed the smallest p-value and highest Kolmogorov-Smirnov distance of all other methods. Additionally, the real network distribution is shifted to lower values, which in this case means that the expression profiles of the real network

were better fitted. A p-value several orders of magnitude lower than other methods provides a reliable statistical measurement of the performance of the discretization procedure.

Regulator Binding Site Multiplicity in the Target Promoter Region

Given that this work uses a regulatory network derived from transcription factor binding site motifs, and that their location and number in the genome is known, we sought a mechanistic relationship between the two. Specifically, we investigated a model that would explain the delays that were found from different rates of accumulation of a transcription factor in a promoter region based on the strength and number of motifs for that factor.

On one hand, we found a striking linear relationship correlating motif instance numbers with the inferred delay of a target gene (Figure 4). This relationship disappeared in a control experiment done by taking randomized regulator profiles and estimating the delay using the multi-step functions method, while keeping the true target and the same number of binding sites. This is consistent with a mechanistic model of activation dependent on the rate of accumulation of a transcription factor on the promoter region of a target gene. The control suggests that the multiplicity of binding sites of a particular regulator may be responsible for the increased delay.

On the other hand, we observed an inverse relationship between the information content, computed by correcting for background frequencies (15), of a sequence motif and the number of instances of that motif in promoter regions of its target genes (Figure 4). Indeed, a motif with higher information content is going to be, by definition, more rare and have fewer occurrences in the genome by chance. Therefore, a weaker promoter, defined here as a transcription factor that needs a higher level of concentration to be able to activate the target gene, should have lower information content and a higher number of binding sites.

Taken together, as regulators with a higher number of binding sites tend to have motifs with smaller information content, our results indicate that weaker promoters exhibit longer delays, consistent with a mechanistic model of protein accumulation dependent on binding site number and degeneracy.

Intrinsic Regulator and Target Functional Delay

For all GO terms associated with the regulators of the network, we determined the average target delay for the genes that are regulated by that given transcription factor and share the same annotation term (Figure 5). Genes with specific biological functions seem to exhibit characteristic average delays. We investigated if a given transcription factor has a characteristic delay distribution, as would be expected if the strength of a promoter is correlated with the delay

of its target genes. Assuming that the distribution of delays for each regulator is an exponential distribution, supported by the fact that the time until induction of each edge is a stochastic process, independent of each other and continuous in time, we fitted an exponential distribution function (parameter λ) to each. A z-score is computed using 1000 random degree preserving networks. The list of transcription factors with a z-score more than 5 standard deviations away (fastest and slowest) from the mean distribution is shown (Figure 5). For both sets we performed GO term enrichment analysis for their target genes using GOrilla (16).

Both sets have unique enriched GO terms (Table S1 and S2), supporting a distinct functional classification, but also share some more general terms. Regulation of a given process appears to be achieved by combining different transcription factors taking into account their particular dynamics. Nevertheless, there are obvious functional differences between both sets of transcription factors.

Edge Type Estimation Validation and Regulation Co-localization

It has been surprisingly difficult to infer computationally whether a regulator is an activator or repressor, perhaps due to the dual nature and context-dependent activity of many regulators. The approach presented here enables classification of each regulator's activity in a target-specific way based on the pairs of transitions in their expression profiles. In order to assess the quality of our predictions we performed a computational validation using the ImaGO annotation database (17), which contains tissue-specific expression annotation during the first 12 hours of the embryonic development for 5979 *Drosophila* genes, overlapping with 2934 genes in the true regulatory network. We compared the interaction types predicted by multi-step functions to those predicted by ImaGO, assuming activators would be physically co-localized with the target gene and that repressors would not. We assessed the predictive power of multi-step functions against the Pearson correlation using ImaGO co-localization for the true regulatory network and also for randomized pairs of genes (Figure 6). We found that for pairs of genes that displayed higher than 0.7 correlation in their expression profiles, the Pearson correlation was a slightest better predictor of co-localization. However, for genes that showed between 0.0 and 0.5 multi-step functions strongly outperformed the Pearson correlation method.

Overall, multi-step functions displayed similar accuracy to the Pearson correlation (Table S3) in the true network and an increase compared to the random network (3% higher). Thus, compressing the information contained in the expression profiles to a much smaller number of parameters describing the discretization still holds high co-localization predictive power, as the

discretization approach can pick up subtle regulatory relationships even when the entire expression profiles are not correlated.

Multiple Regulators and Their Combined Effect on Target Expression

While we have thus far focused on pair-wise regulatory relationships, the typical target gene has multiple regulators controlling its expression, whose interplay is ultimately responsible for the target's expression. Moreover, the relationship between multiple regulators can include complex combinatorial effects such as changing the behavior of a given regulator's sign depending on the context of other regulators, as has been reported for multiple regulators (18).

We addressed this issue by testing whether multiple regulators have coherent profiles that would explain the target expression profile. We present an example target gene with a high number of regulators (Figure 7). The target gene has an ON state consistent with the concurrent expression of its activating targets. As the activating regulators are turned off, repressors start being expressed and eventually the target gene expression is suppressed. In this example, the target's expression profile seems to be the sum of the regulators' expression states, accounting the type of interaction. It should also be noted that *Hr46*, *en*, *run*, *repo* and *snail* display faster target induction independent of the activating or repressor role while *ap* and *retn* show slower induction, consistent with reported results (Figure 5).

Taking advantage of the multi-step functions discretization nature we quantified the independence of the targets and the sum of the regulators expression profiles using the mutual information metric for discrete random variables (see Methods). Lower mutual information values indicate independent variables, which in this context imply that known regulators do not perfectly describe the expression of the target gene. The distribution of the mutual information metric for all the targets, which are not themselves regulators, shows a shift to higher values than randomized degree preserving networks (Figure 7). Given that regulators can better explain overall target expression profiles this suggests that in this dataset there is a lower degree of complexity attributed to the multiple regulators interactions implying that in this dataset the effect of multiple regulators have a tendency to be linearly additive nature. It also makes an argument for the quality of the regulatory network, as the true regulators are able to better explain the targets expression profiles.

Dynamic Network Motif Analysis

Network topology motifs have been proposed as the building blocks responsible for controlling biological functions(19). The possible dynamics of simple network motifs have been worked out,

and demonstrated to perform these roles both theoretically and experimentally (20; 21), at least on a small scale. To further test whether such motifs may play temporal roles in assuring expression of target genes at precise intervals perhaps by filtering biological noise, we studied the temporal dynamics of two of the simplest network motifs, the transcriptional cascade (TC) and the feed-forward loop (FFL).

Instances of the Transcriptional Cascade motif can be separated in activators (types TCa and TCd) and repressors (type TCb and TCc) of the downstream target gene. We studied the overall dynamic properties of transcription cascades using the delay estimates for each edge and found temporally distinct classes averaging a different time until activation or repression of the downstream target gene (Figure 8). Transcription cascade instances where the first edge acts as a repressor show a longer overall delay than when the first edge is an activator (Table S4). Surprisingly, both types appear to have two different temporal forms as type TCb exhibits a faster repression than type TCc and type TCa has a faster activation than TCd. One would expect the overall delay of this motif to be the sum of the average delay of the two types of edges, activators and repressors, giving rise to three particular response times, (assuming different average delays for both types of edges) where TCb and TCc should exhibit the same delay distribution. Instead, we observed an empirical coupling of delays which produces only two different overall dynamics for each type of transcription cascade. The ratio of both delays in each type of transcriptional cascade explains this behavior. Both TCb and TCc activating edges are faster than the corresponding repressing edge. However, the TCc motif has a higher log-ratio between edges that increases the total delay of the motif. These two different populations of motifs, both repressing the downstream target gene, can be chosen to achieve a required time until expression. In addition, for the activating motifs, the second edge is slower regardless of its type, thus allowing this motif to possibly filter out noisy expression of the first regulator (X), while still achieving an even longer difference in the time until expression compared to the repressing motif types. This behavior hints at the possibility that evolution may select the overall temporal timing of regulation even while maintaining the same function.

The feed-forward loop has eight types that can be classified in two major subtypes, coherent and incoherent, based on whether the sign of the direct path is the same as the sign of the overall indirect path (9). In *E. coli* and yeast the coherent type-1 FFL (C1-FFL) and the incoherent type-1 FFL (I1-FLL) occur much more frequently. We enumerated all instances of the feed-forward loops in the *Drosophila* regulatory network and estimated the overall delay of the motif, from the first node (X) to the downstream target gene (Z). Taking a conservative approach we only analyzed instances where the delay of the direct path is equal to the indirect path,

serving as an additional check for meaningful biological motif instances (Figure 8). One striking observation is that in *Drosophila* there is an overwhelming enrichment of the number of coherent-type feed-forward loops and depletion of all incoherent types, including the I1-FFL. The overall dynamic of coherent-type feed-forward loops, in light of the time to full expression of the target gene (Z) can be summarized by a Poisson distribution, which arises from the sum of exponential distributions of the two edges on the indirect path. Using the mean parameter (λ) the overall timing of each type of motif can be compared.

Coherent feed-forward loops will regulate the target gene in an analogous way to the transcriptional cascades as the indirect path does not contradict the direct path logic, but these motifs will have a shorter time until expression than that of transcriptional cascades, as has been previously described (9). For example, C1-FFL and TCa are composed of only activating edges but the feed-forward loop is approximately 2 hours faster than the transcription cascade in our data. The most abundant feed-forward loop in the *Drosophila* network is the C2-FFL, the absolute opposite of the I1-FFL in terms of edge type, which exhibits a longer delay and also shows the biggest difference in delay distributions over the random model. This motif can be paired with C3-FFL, as both of them will ultimately repress the transcription of the target gene upon the activation of the first node. They have a statistically significant different mean time until expression (Kolmogorov-Smirnov test, $p < 0.05$, Table S5 and S6) of about one hour. The activating pair C1-FFL/C4-FFL shows different times until expression of about 2 hours. Target GO term enrichment analysis (Table S7) revealed an enrichment of both faster and slower motifs to the general function of organ development, as there are some biological processes that require a mixture of these effects. Faster types were involved in signal transduction and cell communication, and slower types appear to be associated with metabolic processes.

Discussion

In this paper, we propose a novel unsupervised method to discretize time-course expression profiles that we show to be scalable, robust and does not assume a particular distribution or patterning of the data. Combined with a simple procedure of estimating causal effects for pairs of regulators and targets, it is able to determine a delay between expression profiles displaying an higher accuracy then other methods. In addition, the discretization procedure could be further adjusted by increasing penalties for adding state changes, but in this work, we used a linear penalty model.

Furthermore, using both the predicted regulatory network and the time-course expression profiles we predict that *engrailed* and *Kruppel* will regulate *couch potato* expression. This is

supported by the fact that *couch potato* changes to an ON state during the ON state of both these regulators, showing that an offset between the activation of the regulator and its target gene occurs. Nevertheless this offset could have been bigger, especially if the regulator had a small interval of activation or if the target had a high activating threshold for the concentration of the regulator. In this case, the activation of the target could happen after the deactivation of the regulators.

Drosophila displays a specific non-random delay distribution not present in yeast. The similarity between the predicted network and the Chip IP experimental determined network delay distribution is remarkable, attesting to the accuracy of both the predicted network and the delay estimation method. We argue that the fundamental aspect of temporal regulation at a single edge has been overlooked, since the main model organisms were regulation is studied at the network motif level, yeast and *E. coli*, do not seem to exhibit it. This result clearly shows that a given species may be able to fine-tune its overall response rates by selective changes in the temporal dynamics of any given regulatory interaction, either by adjusting the production rates of a given gene or by tweaking other regulatory mechanisms.

An intuitive explanation for the increased overall delay comes from classical approaches of modeling regulator and target gene interactions in both *E. coli* and yeast. Describing the concentration of the target gene over time with two parameters, production and destruction rates respectively, where destruction is the sum of the dilution and degradation rates, will hold a concentration that will approximate asymptotically the ratio of production and destruction at steady state. In this scenario, the half-life of the target gene will be equal to the response time, defined as the time to reach halfway between initial and final concentration levels. For proteins not actively degraded, the dilution rate governs the destruction rate, which for yeast is one cell-generation, resulting from cell growth and ultimately determining a uniform response time for most genes by dampening changes in promoter strength. In *Drosophila* on the contrary, although there are cellular divisions, the volume of the embryo increases orders of magnitude more slowly so it is not possible to control response time by dilution. In order to avoid producing excess quantities of a given *Drosophila* gene, promoters have to be generally weaker than in yeast, leading to increased response times that manifest as delays. Furthermore, several other factors may play a role in genetic regulation in higher organisms, as discussed above.

We find a correlation between the number of regulator binding site motifs on the promoter region of a given target and its induction delay, suggesting that weaker promoters lead to increased delays. However, we did not find a correlation with motif information content or an increased correlation incorporating both the multiplicity of binding sites and information content.

A possible mechanism that would explain this behavior is the requirement of a large fraction of occupied binding sites until expression is activated or repressed. Such a mechanism would compensate noisy regulators for genes that require increased delays. In addition, this would provide a fast evolutionary mechanism to control expression latency by tuning the number of motif binding sites. From the regulator point of view, we showed that there is an intrinsic dynamic behavior for each transcription factor with some regulators acting mostly with fast or slow dynamics, and there is a different functional enrichment for the targets of fast and slow regulators.

Although there is currently no standard for assigning the type of interaction a regulator has with its target, we could derive an intuitive classification from our method and validate it against the ImaGO annotation. Expression profiles do not show extremely strong predictive power of co-localization in the ImaGO gold standard (Table S3), as only profiles with high correlation exhibit high accuracy. However, given that the probability of finding two co-localized tissues in ImaGO is low, the accuracy of the Pearson method for negative correlations is high, though not informative. Multi-step functions performed approximately as well as Pearson correlation predicting co-localization.

ImaGO has several caveats, as it only has annotations for a small set of genes expressed in the first 12 hours of embryo development, which may lead to ascertainment bias if most of the repressing interactions take place after this. In addition, this annotation does not intrinsically define time so it is possible that, although there exists a true regulatory activating interaction between a regulator and a target expressed with a true delay, the target gene will be annotated with a child tissue, developed from the ancestor where the regulator was expressed, thus being wrongly classified as not co-localized. Furthermore, this annotation does not contain information on expression gradients.

For instance, one example of this behavior is the expression of *bicoid* and *hunchback* gradients (22). *bicoid* is expressed at the anterior tip of the oocyte during early embryonic stages producing a gradient in the anterior-posterior axis and is known to activate the expression of *hunchback* although they only exhibit a small physical overlap. Our method classifies *bicoid* as an activator of *hunchback* while ImaGO does not. For repressing interactions, it is also possible that expression of regulator and target overlap in the same tissue for a small time interval, as the concentration of the regulator increases to repress the transcription of the target. Thus, comparing expression profiles to ImaGO annotation might hold a reduced accuracy for both co-localized and not co-localized genes.

Our approach does not take into consideration the complex combinatorial effects that a particularly high number of regulators might exert on a single target. Nevertheless, we showed that this dataset exhibits a lower overall combinatorial nature. A large percentage of the targets in the network only have one known regulator. This leads to biased distribution estimation, as there is a smaller chance that the regulator expression profile will be able to completely explain all the time points of the target profile. Thus, the mutual information distribution of the true complete network should be higher than the estimated. Decreased complexity does indeed mitigate errors on the edge type estimation in this work. However, our approach does not take into account other sources of regulatory complexity as the typical latency that a combination of transcription factors might exhibit, nor any other post-transcriptional regulatory mechanism

Surprisingly both the transcriptional cascade and the feed-forward loop exhibit two different distributions of time until expression of the last target gene for both activating and repressing motifs. This implies that particular motif types may be used because of not only the robustness and noise filtering properties already described but also to meet a particular delay during development or even in adult life. This might explain why some biological modules are composed of what could be classified as non-optimal topologies in light of network structure alone, but may be optimal when one considers temporal aspects. other criteria.

Taken together, these results seem to suggest that the *Drosophila* regulatory network possess a different underlying architecture that accounts for the delay in each individual edge, either by tuning each edge distribution to stay in comparable intervals of time to achieve synchronization, or by choosing from a plethora of network motifs that will hold particular temporal dynamics.

Materials and Methods

Microarray Data Normalization

For both *Drosophila* and yeast, time-course expression profiles where normalized by subtracting the mean and dividing by the standard deviation across all time-points of the gene-wise expression profile. Technical replicated expression profiles were merged by averaging each time-point. Any incomplete expression profile was discarded from the analysis.

Multi-step Functions

We developed a robust method to find gene expression states from microarray time-course data. Assuming that a given gene has only two types of states, ON and OFF, and that each state has a minimum length of two time-points, we define all the possible models (M_i) by

enumerating all combinations of the two types of states up to a maximum number of state changes between these state types (S_{max}). Every time-point is then effectively assigned to a state and there will be at most S_{max} state changes in the time course classification. Each model is built from a number of parameters (k) proportional to the number of state changes. In order to chose the gene model that best discretizes the expression profile we score each of them using the following equation:

$$M_{score} = \log\left(\frac{\sum_{j=1}^{|S|}\sum_{i=1}^{|n_s|}\left(x_i^j - \bar{x}^j\right)^2}{n-k}\right) + \frac{n+k}{n-k-2}$$

where S is the state list of the gene model, x is the time-course expression profile, n is the length of x and k is the number of parameters used to construct the gene model. This score represents the root mean square deviation for each time-point (x_i) to the average expression of the state where it has been classified (j), balanced by the number of model parameters using a corrected Akaike information criterion (23). Choosing the model with the minimum M_{score} holds the one that best describes the expression profile of that particular gene without overfitting the data. Additionally, a gene model without state changes is also scored, allowing for basal expression profiles.

We set the maximum number of state changes (S_{max}) high enough to capture the real gene state changes and still being computationally feasible, as the number of the possible models for a given expression profile increases with the number of time-points and state changes.

True Predicted Network and ChIP-chip Experimental Network

In this work, the predicted *Drosophila* regulatory network by sequence conservation from closely related species was used at a 60% confidence as the *de facto* network (8). For a subset of this network, validation from chromatin immunoprecipitation was also published and analyzed and is used in the paper as the experimental validated network. For yeast, the experimental derived regulatory network from Lee at al. (14) was used at a p-value of 0.005.

Computing Delay and Type of Interaction between Genes

From the discretization of the gene expression profiles, we assumed that a state change in the regulator gene would be responsible for the next state change in the target gene. We find all the pairs of related state changes, compute the average delay between state changes and assign it

as the delay for a pair of expression profiles. The regulator state change is always paired with the closest downstream target gene state change. If no state change is found in the target after the regulator state change the edge is discarded.

This pairing of state changes also allows us to evaluate if a given regulator is an activator or a repressor of the target gene. For each interaction pair we determine if the state change is consistent with an activator or repressor behavior. The results are very consistent but whenever there is a conflict, we use a majority rule to compute the overall type of interaction, and discard any tied result.

Cross-Correlation Function and Sliding Pearson and Spearman correlations

Correlations were computed only for a sliding target expression profile that represent a true causal relationship to the regulator up to a cut-off number of time-points, meaning that a delay between a regulator and a target will always be computed having a positive value. The number of time-points used decreases while increasing the offset. The offset with the highest absolute correlation value for a given regulator/target expression profile is used and the delay is computed by the difference between the first time-point label of the regulator and the target offset label.

Gene Ontology Term Enrichment

The GO term enrichment analysis was done using the GOrilla website tool (16). Comparison were always performed with a query set of interest against all the genes in the microarray, using the two unranked list of genes options for *Drosophila* and a p-value threshold of 0.001.

ImaGO

ImaGO is a *Drosophila* anatomy ontology organized in a direct acyclic graph (DAG) and modeled according to GO. Manual curators annotated *in situ* expression patterns of a large set of genes (n=5979) during embryonic development using this controlled vocabulary. The vocabulary includes definitions that have a temporal component linked to the developmental stages, but does not possess an implicit way of accurately relating ontology terms with developmental time. Furthermore, each gene is annotated with a set of broad terms corresponding to six bins of the embryonic developmental stages.

Certain ImaGO terms were handled specially given that they are too general and do not describe a particular tissue. A example is the term ‘ubiquitous’. Whenever at least one of the genes in the pair is annotated with this term then they are always assigned to be spatially co-

localized. We found that about 38% of the 17.87 x 10^6 possible pairs are spatially co-localized in at least one tissue. We defined a pair of genes to be spatially co-localized if both of them are annotated in ImaGO and they share at least one non-general term.

Mutual Information

For a given target, all the binary discretized regulator profiles were added into a vector. Gene-wise activating regulators were described with 0 and 1 and repressing regulators with 0 and -1 for OFF and ON states respectively. The final regulator vector varies between –n an n (n is the number of target's regulators). Mutual information is computed using the regulators vector sum and the target discretized profile using a discrete joint probability distribution function.

Random Networks and Network Motifs

Degree preserving networks were constructed by switching node labels by any gene present in the time-course microarray dataset while maintaining the original network topology. For the binding site multiplicity control random network, only regulator expression profiles where switched, while the target expression profile and the number of a given original regulator binding sites were maintained.

Network motifs were extracted from the network using the igraph python library (http://cneurocvs.rmki.kfki.hu/igraph/index.html)

Acknowledgements

We thank all the members of the Kellis lab for helpful discussions and feedback on the work, and Joshua Grochow for comments on the manuscript.

Funding. This work was supported by the Portuguese Foundation for Science and Technology Fellowship SFRH / BD / 32967 / 2006 in the context of the PhD Program in Computational Biology of Instituto Gulbenkian de Ciência, sponsored by Fundação Calouste Gulbenkian, Siemens SA, and Fundação para a Ciência e a Tecnologia, Portugal

Competing interests. The authors have declared that no competing interests exist.

References

1. Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, et al. Genome-wide Prediction of Mammalian Enhancers Based on Analysis of Transcription-Factor Binding Affinity. Cell. 2006 Jan ;124(1):47-59.

2. Choi CY, Lee YM, Kim YH, Park T, Jeon BH, Schulz RA, et al. The Homeodomain Transcription Factor NK-4 Acts as either a Transcriptional Activator or Repressor and Interacts with the p300 Coactivator and the Groucho Corepressor. J. Biol. Chem. 1999 Oct 29;274(44):31543-31552.

3. Lam FH, Steger DJ, O/'Shea EK. Chromatin decouples promoter threshold from dynamic range. Nature. 2008 May 8;453(7192):246-250.

4. Baugh LR, Hill AA, Slonim DK, Brown EL, Hunter CP. Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. Development. 2003 Mar 1;130(5):889-900.

5. Li T, White KP. Tissue-Specific Gene Expression and Ecdysone-Regulated Genomic Networks in Drosophila. Developmental Cell. 2003 Jul ;5(1):59-72.

6. Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, et al. A Gene Expression Map for the Euchromatic Genome of Drosophila melanogaster. Science. 2004 Oct 22;306(5696):655-660.

7. Zhu C, Byers KJRP, McCord RP, Shi Z, Berger MF, Newburger DE, et al. High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res. 2009 Apr ;19(4):556-566.

8. Kheradpour P, Stark A, Roy S, Kellis M. Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res. 2007 Dec ;17(12):1919-31.

9. Alon U. Network motifs: theory and experimental approaches. Nat. Rev. Genet. 2007 Jun ;8(6):450-61.

10. Hooper SD, Boue S, Krause R, Jensen LJ, Mason CE, Ghanim M, et al. Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis [Internet]. Mol Syst Biol. 2007 Jan 16;3[cited 2009 Jan 21] Available from: http://dx.doi.org/10.1038/msb4100112

11. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucl. Acids Res. 2007 Jan 12;35(suppl_1):D760-765.

12. Agrawal A, Mittal A. A Dynamic Time-Lagged Correlation based Method to Learn Multi-Time Delay Gene Networks. Proceedings of World Academy of Science, Engineering and Technology. 2005 ;9167–174.

13. Sahoo D, Dill DL, Tibshirani R, Plevritis SK. Extracting binary signals from microarray time-course data. Nucleic Acids Res. 2007 ;35(11):3705-12.

14. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, et al. Transcriptional Regulatory Networks in Saccharomyces cerevisiae. Science. 2002 Oct 25;298(5594):799-804.

15. D'haeseleer P. What are DNA sequence motifs? Nat Biotech. 2006 Apr ;24(4):423-425.

16. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009 ;10(1):48.

17. Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis SE, et al. Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol. 2002 ;3(12):RESEARCH0088.

18. Georlette D, Ahn S, MacAlpine DM, Cheung E, Lewis PW, Beall EL, et al. Genomic profiling and expression studies reveal both positive and negative activities for the Drosophila Myb MuvB/dREAM complex in proliferating cells. Genes Dev. 2007 Nov 15;21(22):2880-2896.

19. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network Motifs: Simple Building Blocks of Complex Networks. Science. 2002 Oct 25;298(5594):824-827.

20. Kalir S, Mangan S, Alon U. A coherent feed-forward loop with a SUM input function prolongs flagella expression in Escherichia coli [Internet]. Mol Syst Biol. 2005 Mar 29;1[cited 2009 Sep 28] Available from: http://dx.doi.org/10.1038/msb4100010

21. Mangan S, Itzkovitz S, Zaslaver A, Alon U. The Incoherent Feed-forward Loop Accelerates the Response-time of the gal System of Escherichia coli. Journal of Molecular Biology. 2006 Mar 10;356(5):1073-1081.

22. Ochoa-Espinosa A, Yucel G, Kaplan L, Pare A, Pura N, Oberstein A, et al. The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. Proceedings of the National Academy of Sciences of the United States of America. 2005 Apr 5;102(14):4960-4965.

23. McQuarrie A, Shumway R, Tsai C. The model selection criterion AICu. Statistics & Probability Letters. 1997 Jun 16;34(3):285-292.

Figure Legends

Figure 1 – ***engrailed*, *Kruppel* and *couch potato* discretized profiles**. A, B and C) Blue lines represent the gene expression profile, red lines the boundaries between different ON/OFF states and green lines the average expression of the gene two states. The expression profiles were normalized to zero mean and one standard deviation. D) *engrailed* regulates *Kruppel* and together both regulate *couch potato*, exhibiting a coherent timing of activation. Numbers represent average state change delays.

Figure 2 – *Drosophila* and yeast network delay distribution. A) Regulatory network (blue), experimentally determined network (green) and 100 degree preserving randomized network (red) delays distributions of *Drosophila* using multi-step functions. B) Yeast experimental determined regulatory network (blue) and 100 degree preserving randomized networks (red). Random networks were binned using the same intervals as the true networks and error bars represent the standard deviation of the delay counts of the particular bin. *Drosophila* exhibits a delay distribution slower than expected, whereas yeast does not.

Figure 3 – Method comparison. A) True network delay distribution (solid lines) and degree preserving randomized network (dashed lines), obtained using several methods. Multi-step functions show a larger difference from the randomized network, while both sliding Pearson and Spearman correlations classify a large fraction of edges with the maximum allowed delay. B) Fitted score (multi-step functions) and correlation coefficients distribution (other methods) for true and randomized networks. For multi-step functions, a left shifted distribution means a better fit, while for the other methods a right shifted distribution corresponds to a higher absolute correlation. C) Kolmogorov-Smirnov test for every method using the regulatory network and randomized networks. Multi-step functions have the lowest p-value and highest Kolmogorov-Smirnov D.

Figure 4 – Regulators information content and target's regulator binding site multiplicity. A) Regulator's sequence motif information content binned by the number of identified instances in the target's 2kb promoter region. Regulators with more binding sites have reduced information content. B) Average delay for all targets, binned by the number of binding sites of

their regulators, the blue line represents the linear regression of the average target delay with the number of regulator binding sites (r^2 = 0.868) and the red line the random regulator (r^2 = 0.002), showing that the average target delay increases with the number of regulator's binding sites. The random regulator network was obtained by keeping the true target and corresponding regulator binding site distribution but computing the delay between a randomized regulator and the target with 30 fold oversampling.

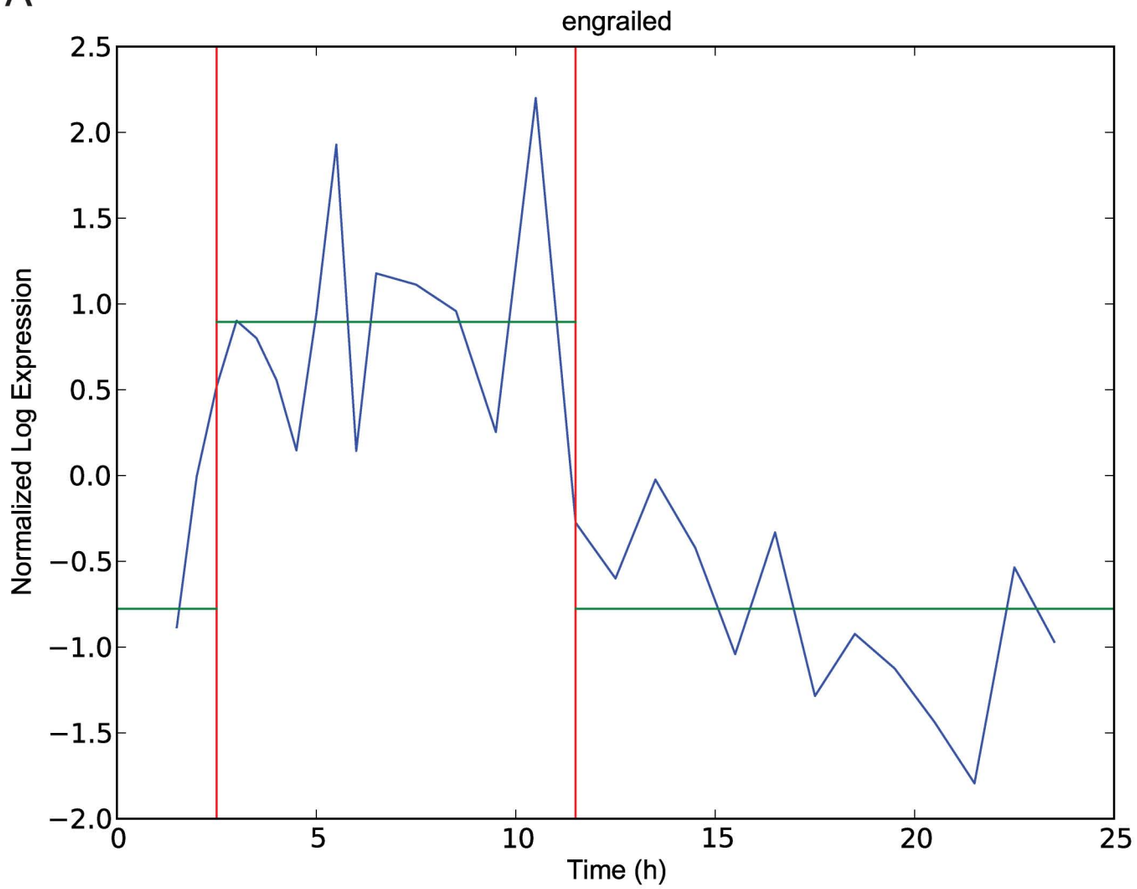
Figure 5 – Target delay by regulator annotated function and intrinsic regulator delay. A) Averaged target genes delay grouped by any shared regulator's GO term (for groups with more than five target members). The same regulators may control process with distinct time scales. B) and C) Faster and slower regulators with more (or less) then 5 standard deviations from the estimated z-score. Z-scores were computed using 300 different degree preserving random networks, taking the estimated exponential parameter λ and its standard deviation.

Figure 6 - Regulation and co-localization in the ImaGO annotation. For the true (blue and cyan) and random network (red and salmon) we present the Pearson correlation distribution (top) and the fraction of correctly predicted co-localized genes in ImaGO (bottom, two genes are considered co-localized if they share at least one non-general ImaGO annotated tissue), binned by Pearson correlation. The multi-step functions discretization method exhibits a higher predicted power for small absolute Pearson correlations for the two types of networks.

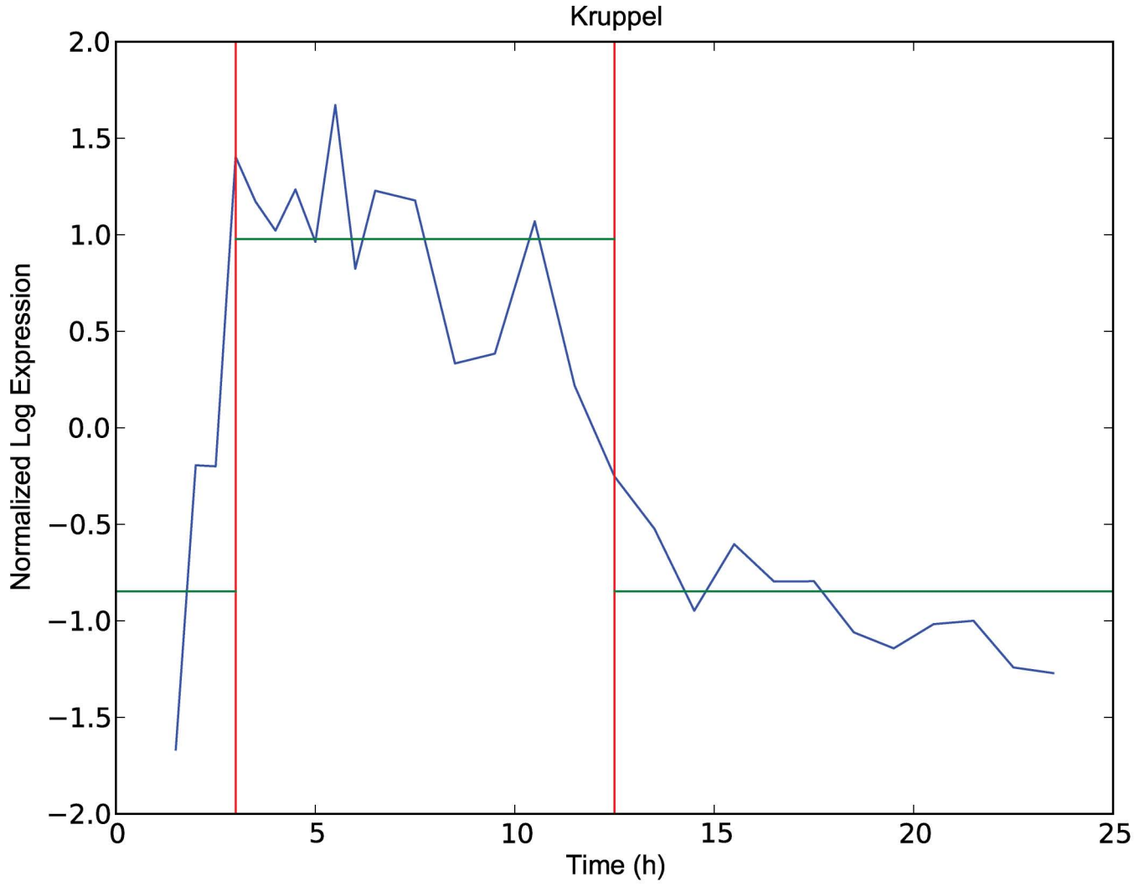
Figure 7 – Multiple regulators profiles describe a target profile. A) Discretized expression profile of CG3225 and all its regulators state changes. Activating regulators are colored green when in the ON state and repressors red. The yellow profile corresponds to a regulator that has no state changes. Although *Dfd* is predicted to regulate this gene its state changes do not display a causal relationship with the target and is thus not considered a true regulator by the multi-step functions method (ON state is colored dark green and OFF state dark red). CG3225 expression profile seem to be explained by all its regulators profiles. B) All target genes (blue) that are not themselves regulators mutual information distribution against 300 degree preserving randomized networks (red). The true distribution is shifted to higher values of mutual information, meaning that target profiles seem to be less independent of the regulator profiles than expected.

Figure 8 – Transcription cascade and feed-forward loops dynamic behavior. A) Average delay until induction of the last gene (Z) in all the instances of three node transcriptional cascades in the regulatory network (left) with mean log-ratio of both delays (right). Positive ratios represent a slower first edge. All distributions are statistically significantly distinct (Table S5 and 6). B) Estimated λ for each type of feed-forward loop when more than 50 instances were found. An instance is only included in the analysis if the sum of the delays of the indirect path is equal to the direct path. Gray bars correspond to the expected λ in an Erdős–Rényi network with the same inclusion criteria. The distributions of all the types of motifs are statistically distinct for the random network and from the other motif types (all Kolmogorov-Smirnov tests, $p < 0.05$).

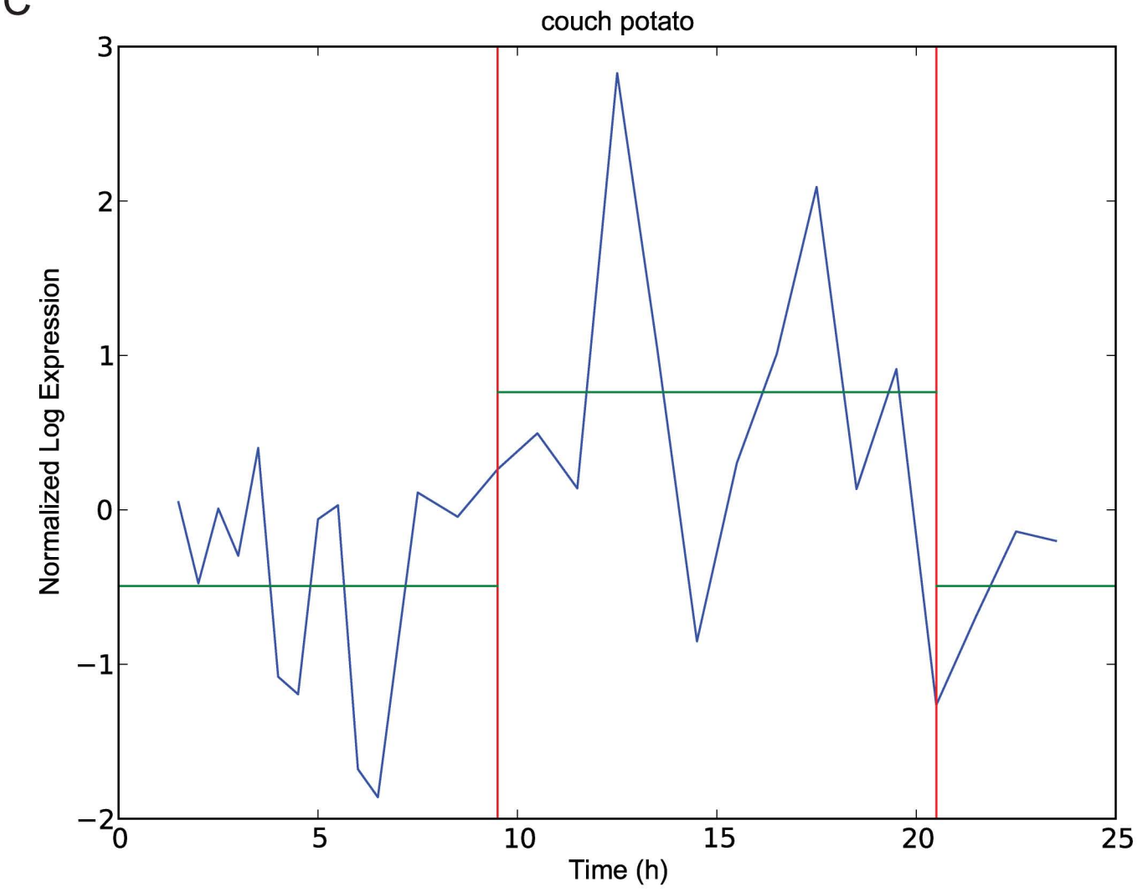
A


engrailed
Normalized Log Expression
Time (h)
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0
5
10
15
20
25

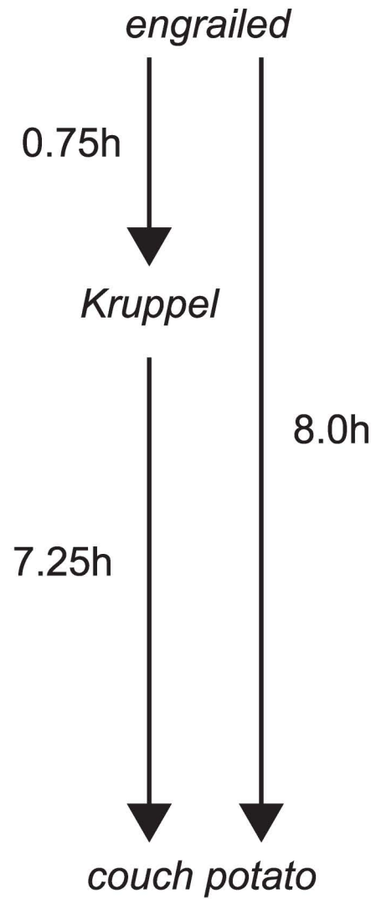
B


Kruppel
Normalized Log Expression
Time (h)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0
5
10
15
20
25

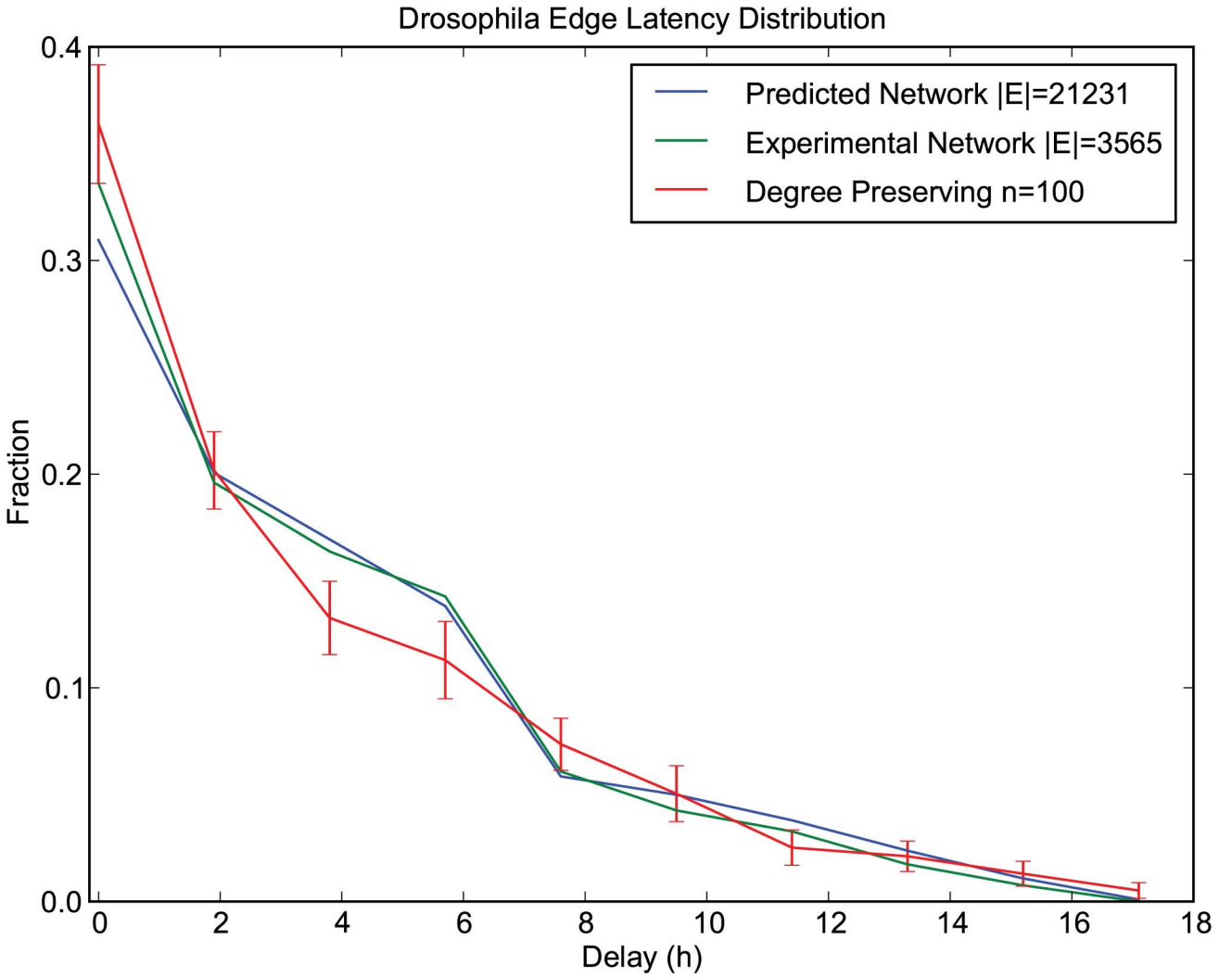
C


couch potato
Normalized Log Expression
Time (h)
3
2
1
0
−1
−2
0
5
10
15
20
25

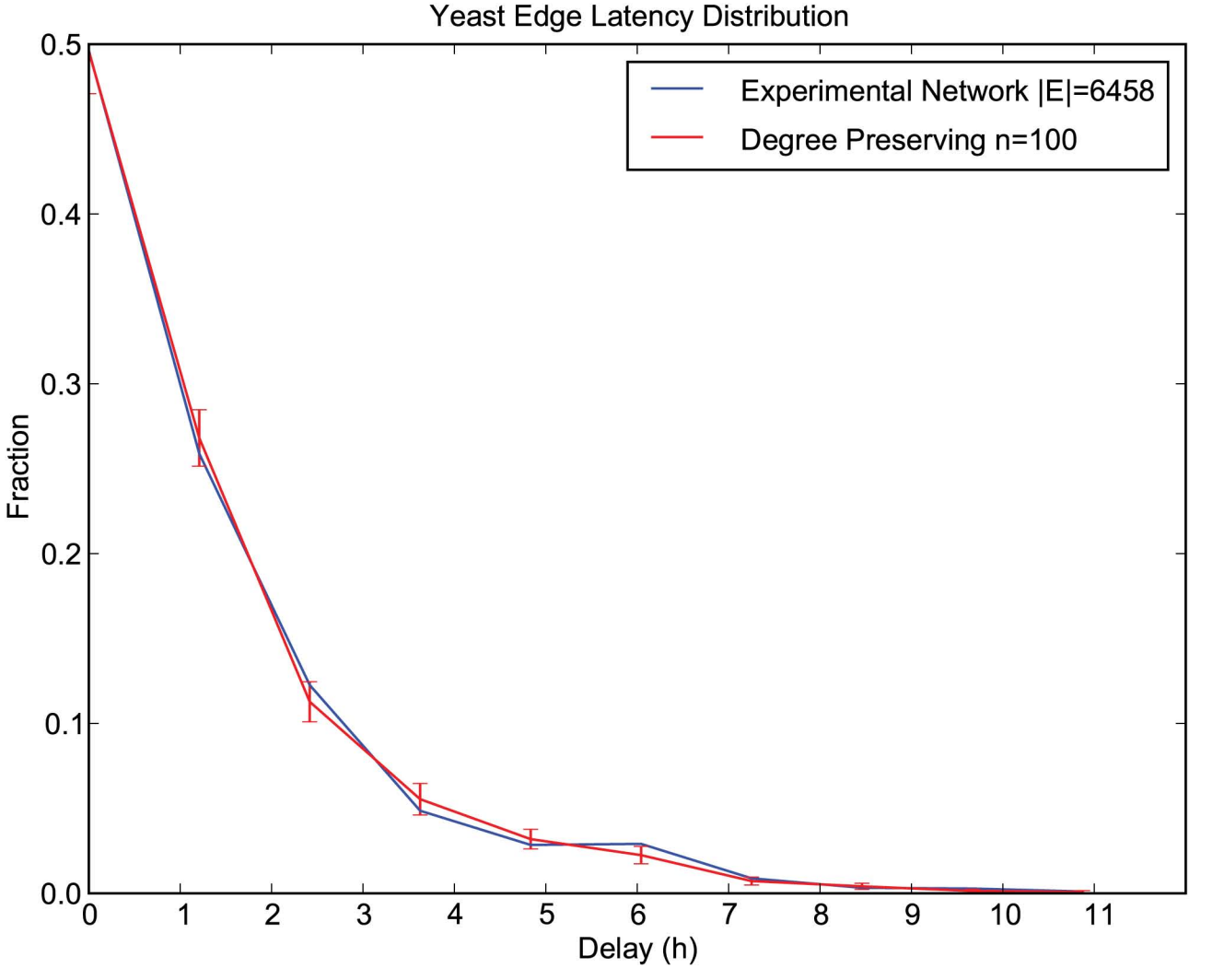
D


engrailed
0.75h
Kruppel
8.0h
7.25h
couch potato

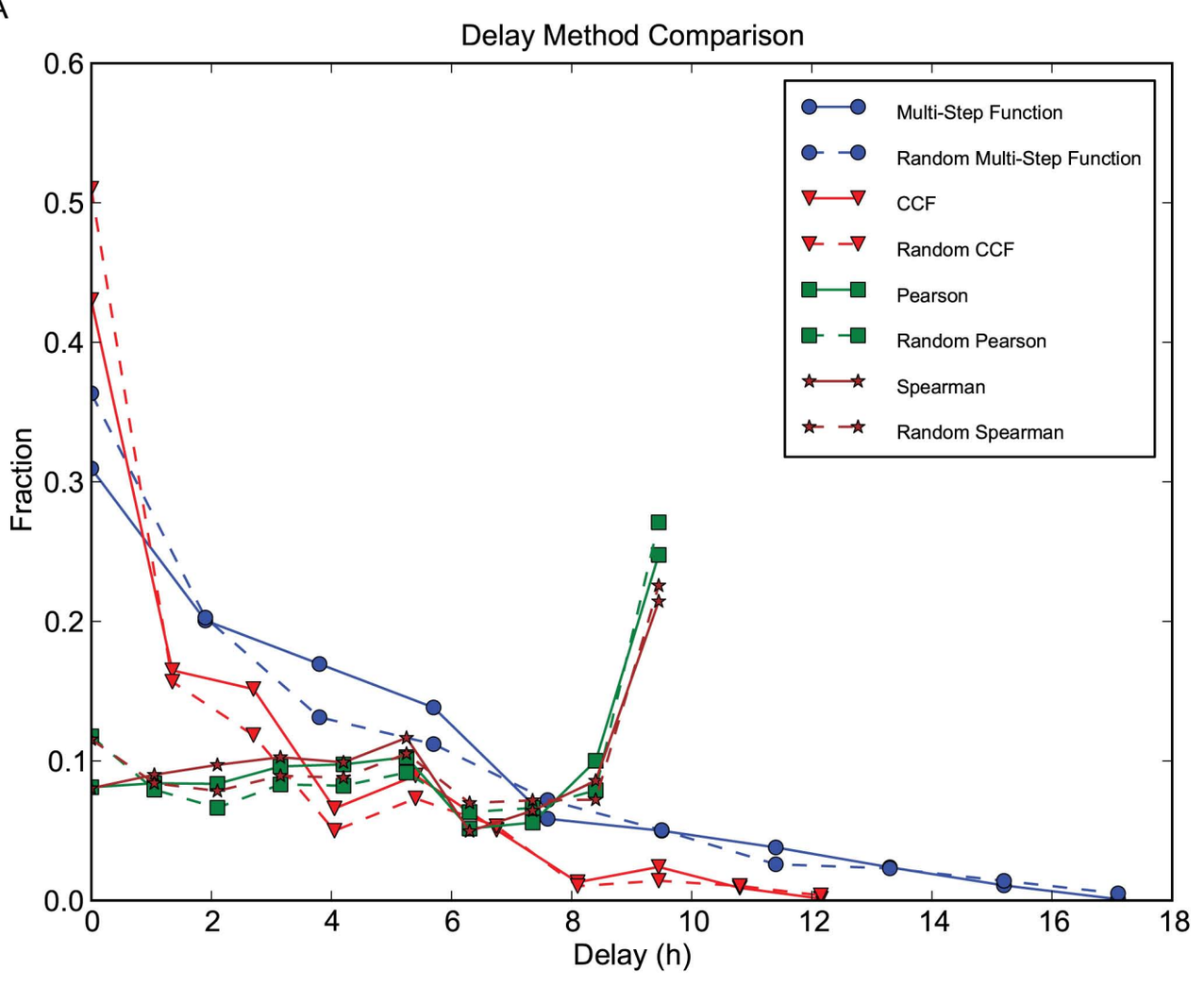
A


Drosophila Edge Latency Distribution
Predicted Network |E|=21231
Experimental Network |E|=3565
Degree Preserving n=100
Fraction
Delay (h)
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10
12
14
16
18

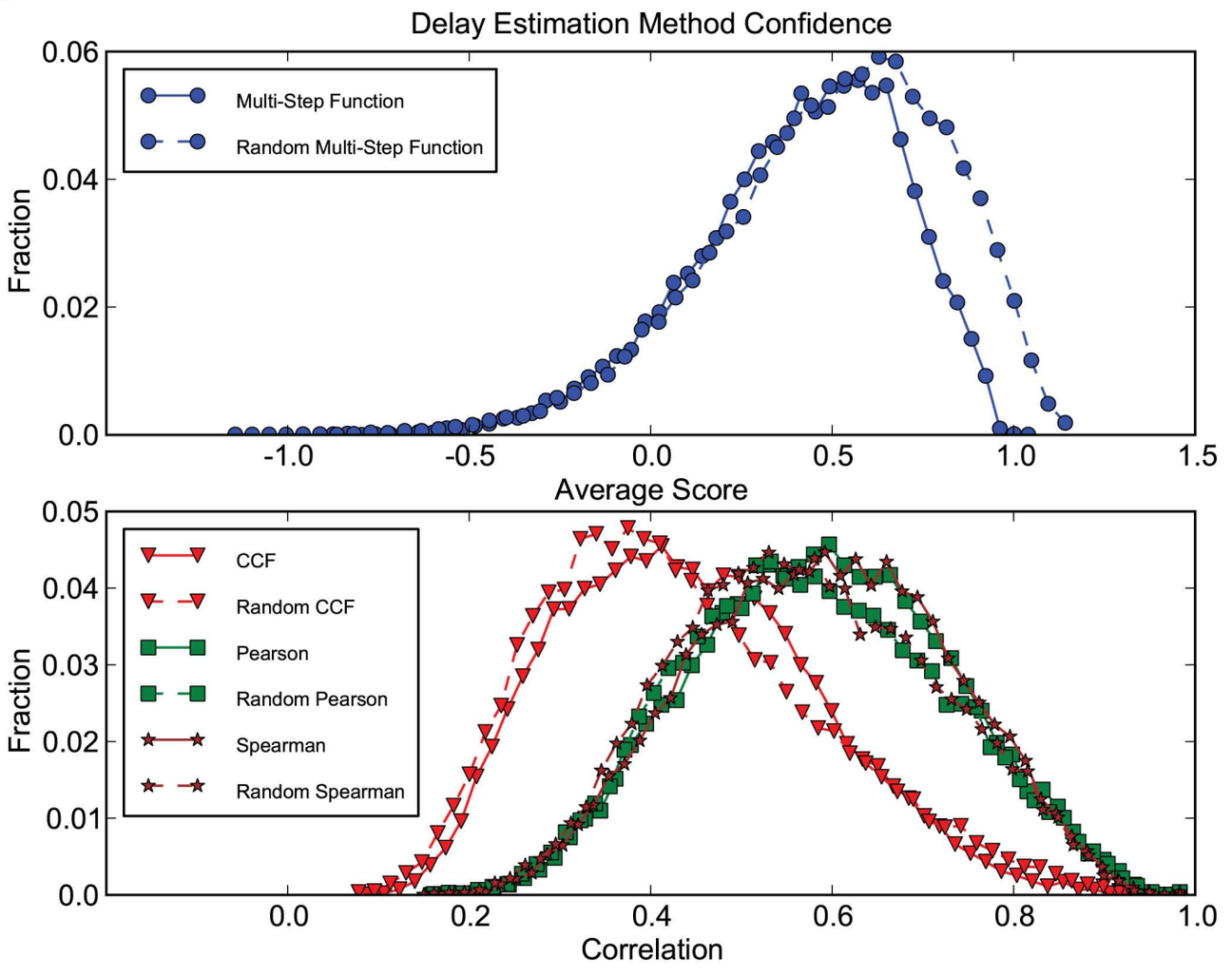
B


Yeast Edge Latency Distribution
Experimental Network |E|=6458
Degree Preserving n=100
Fraction
Delay (h)
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
6
7
8
9
10
11

A



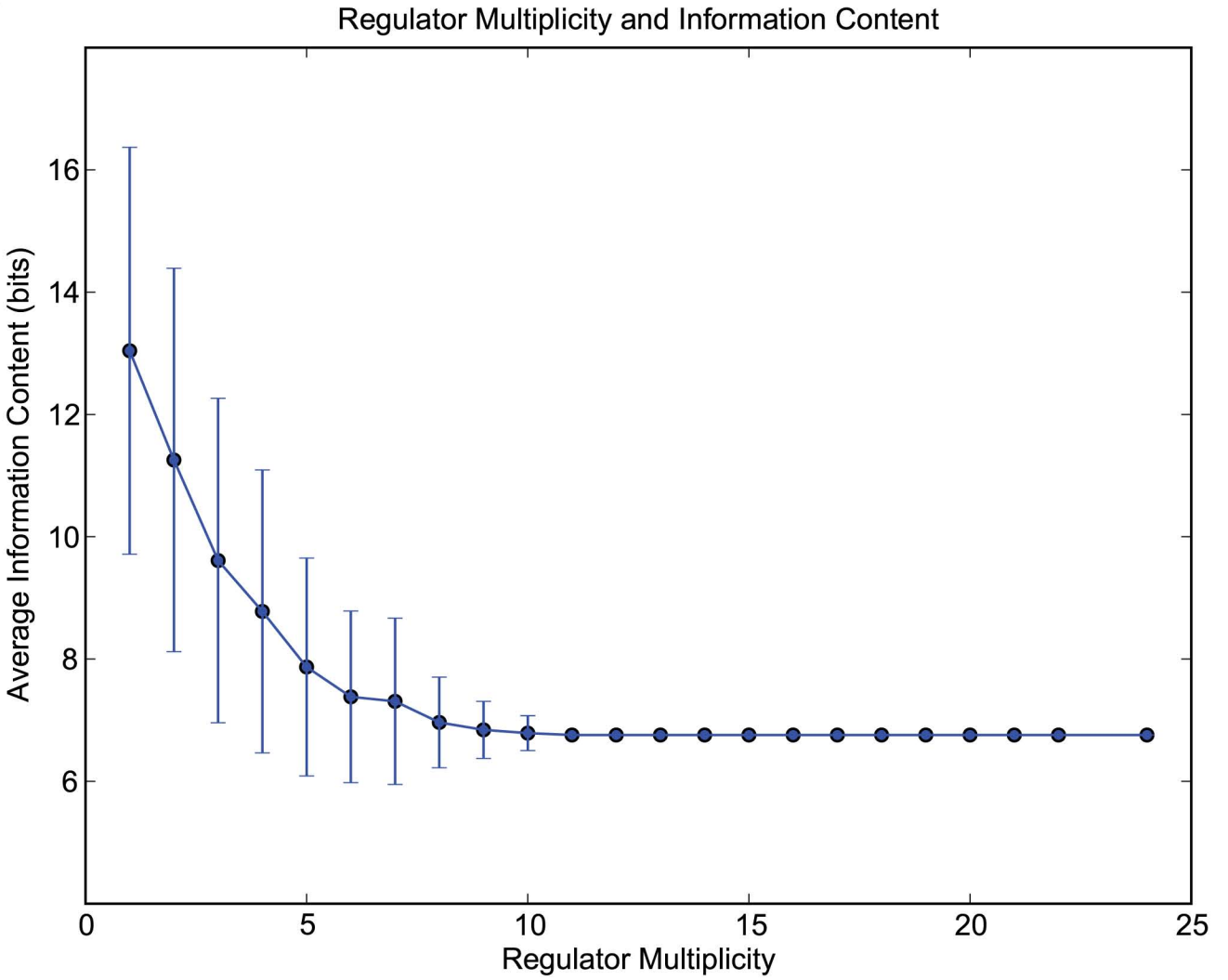
B



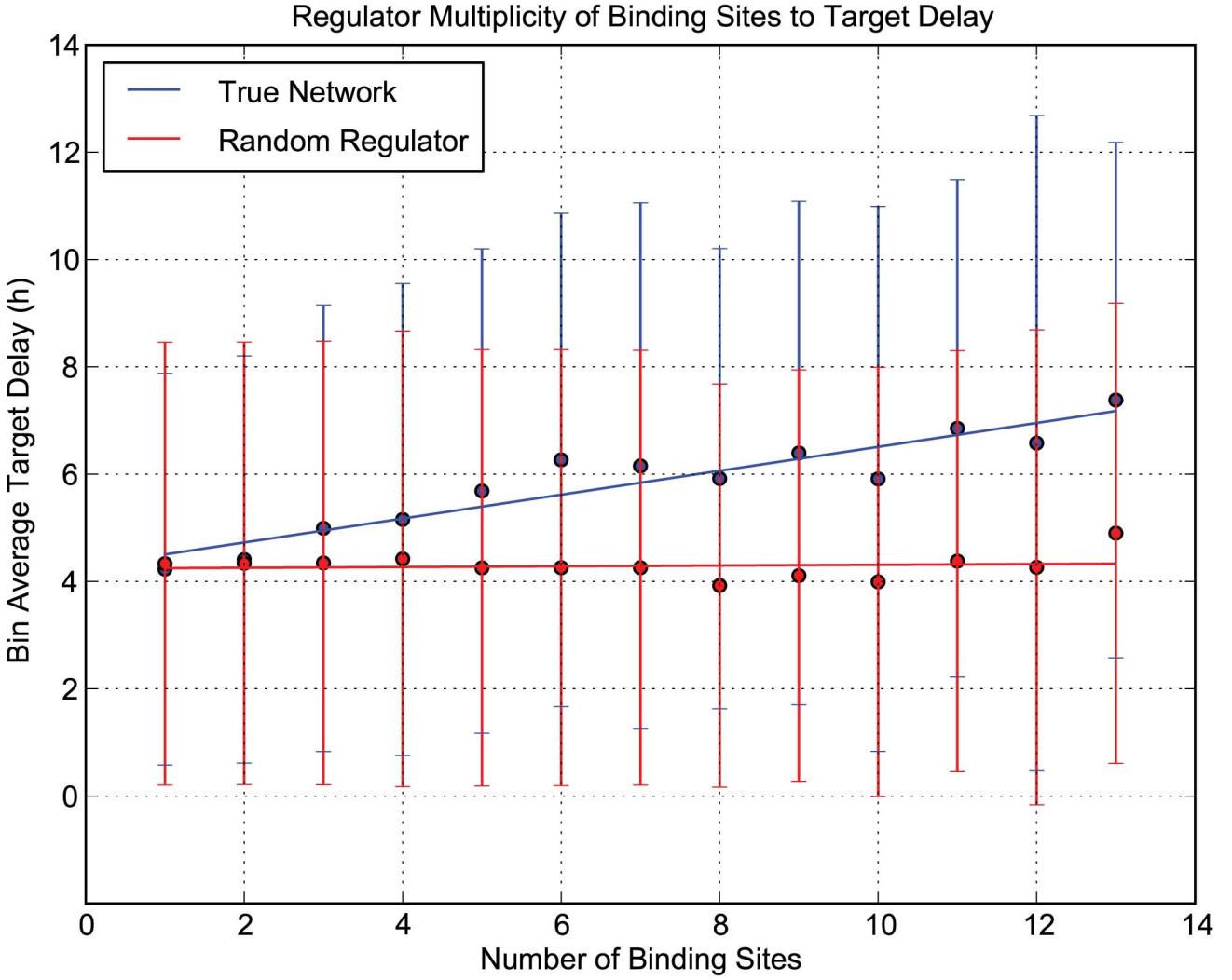
C

Method	Kolmogorov-Smirnov D	p-Value
Multi-Step Function	0.15663684	2.34E-215
CCF	0.05971796	2.23E-43
Pearson	0.04042338	4.13E-20
Spearman	0.05819666	3.16E-41

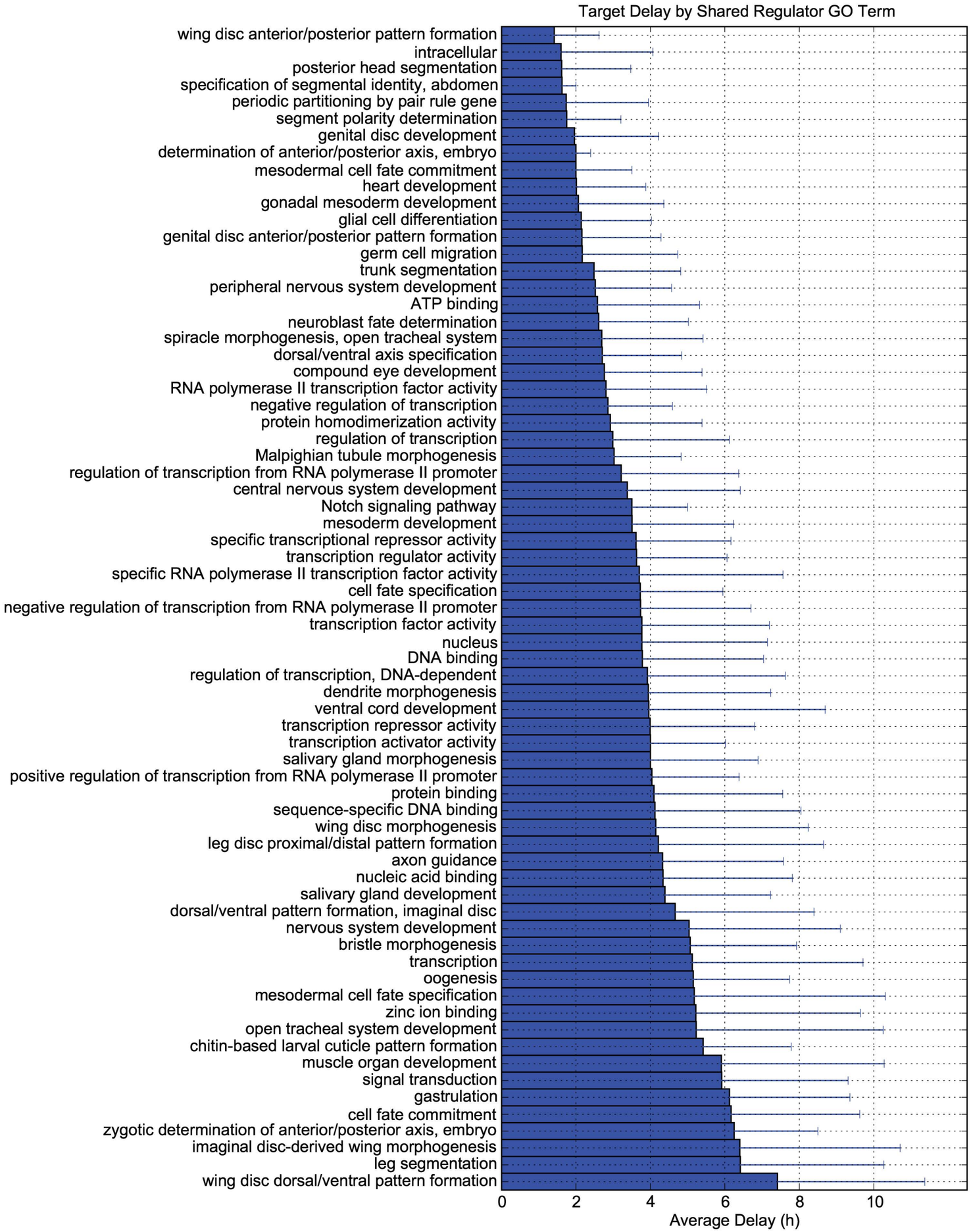
A


Regulator Multiplicity and Information Content
Average Information Content (bits)
16
14
12
10
8
6
0
5
10
15
20
25
Regulator Multiplicity

B


Regulator Multiplicity of Binding Sites to Target Delay
True Network
Random Regulator
Bin Average Target Delay (h)
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
Number of Binding Sites

A



B

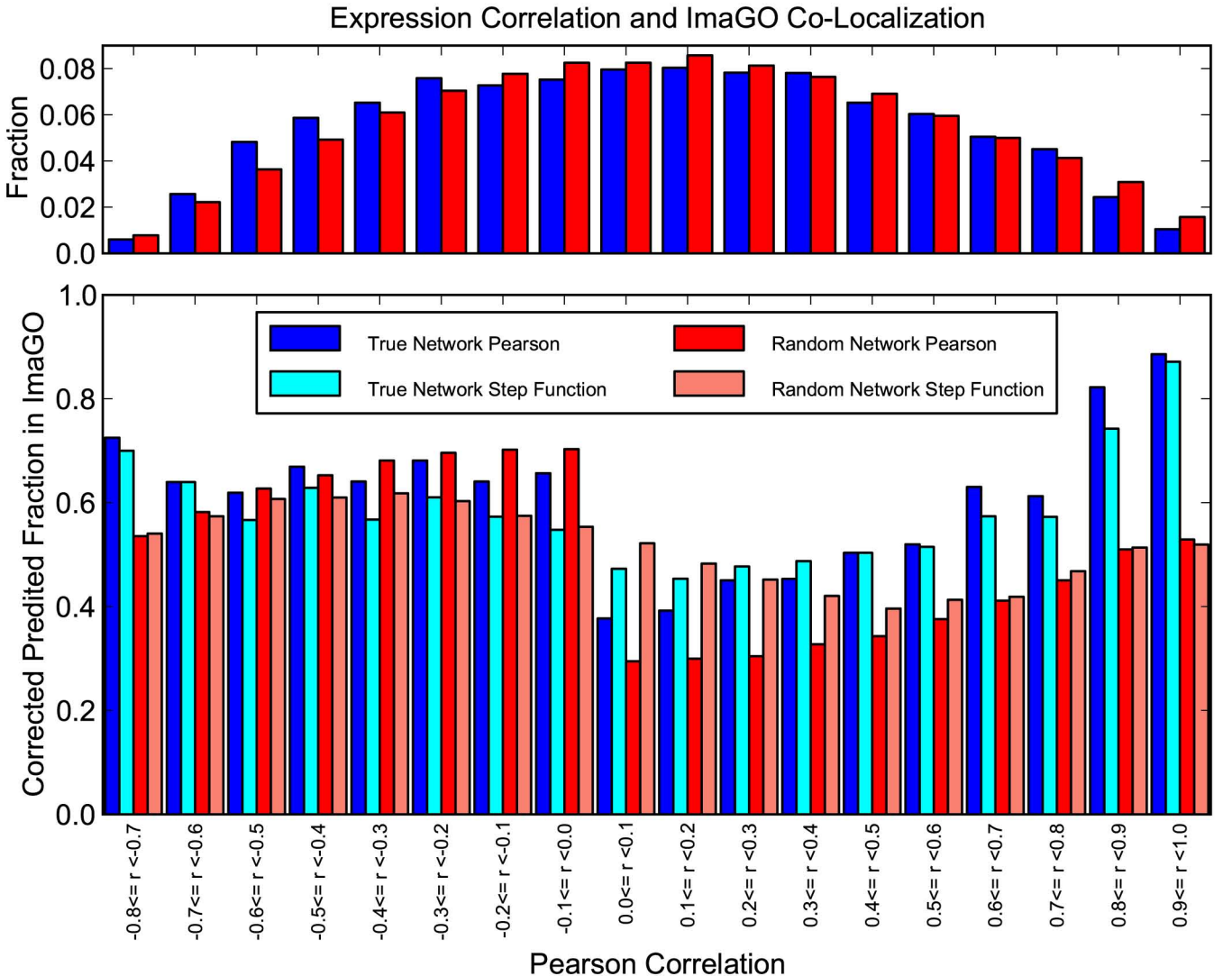
Fast Regulators

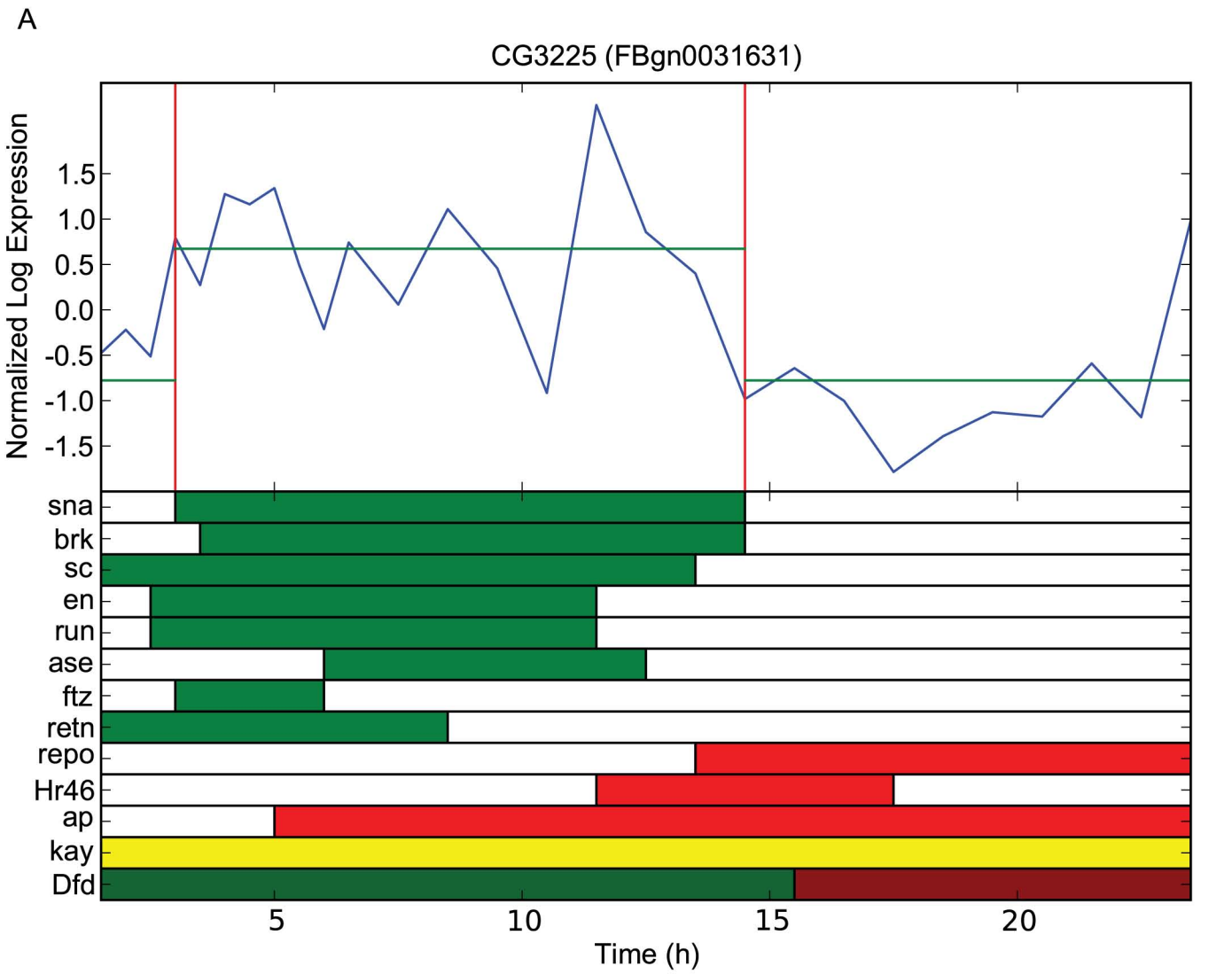
Flybase ID	Gene Symbol	λ	Random Mean	Random Std	z-score
FBgn0000448	Hr46	0.471	0.2263	0.0057	43.03
FBgn0000577	en	0.361	0.2262	0.0033	41.28
FBgn0000439	Dfd	0.387	0.2257	0.0053	30.35
FBgn0003300	run	0.375	0.2258	0.0071	20.90
FBgn0011701	repo	0.291	0.2265	0.0033	19.30
FBgn0001078	ftz-f1	1.600	0.2475	0.0780	17.35
FBgn0000137	ase	0.304	0.2262	0.0053	14.69
FBgn0004170	sc	0.284	0.2265	0.0067	8.57
FBgn0003448	sna	0.267	0.2261	0.0052	7.87
FBgn0000015	Abd-B	0.306	0.2257	0.0102	7.83
FBgn0000576	ems	1.000	0.2523	0.0957	7.81
FBgn0014179	gcm	0.359	0.2279	0.0185	7.08
FBgn0000022	ac	0.413	0.2290	0.0275	6.69
FBgn0004837	Su(H)	0.550	0.2331	0.0474	6.69
FBgn0001325	Kr	0.295	0.2266	0.0127	5.37
FBgn0001981	esg	0.272	0.2272	0.0087	5.21

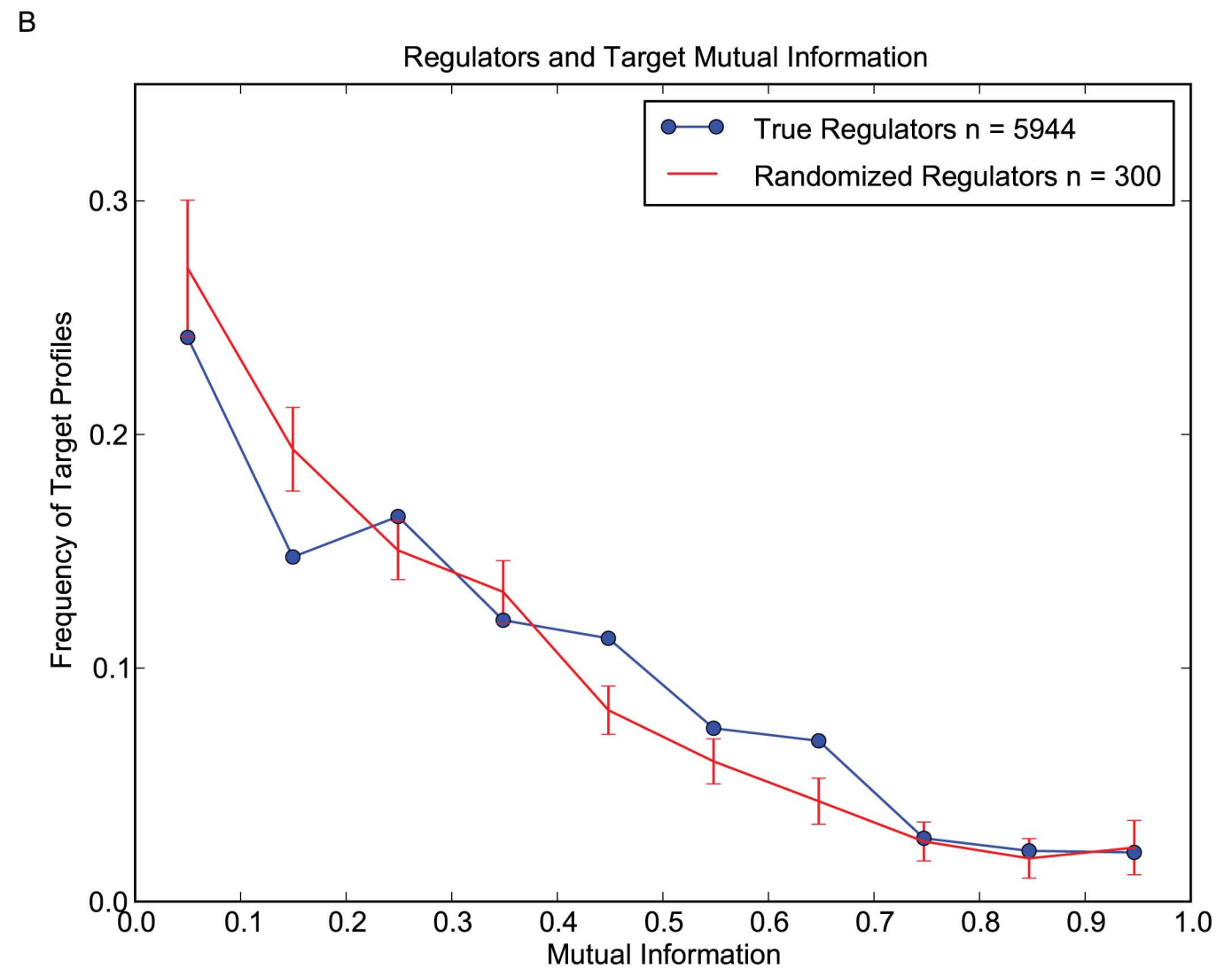
C

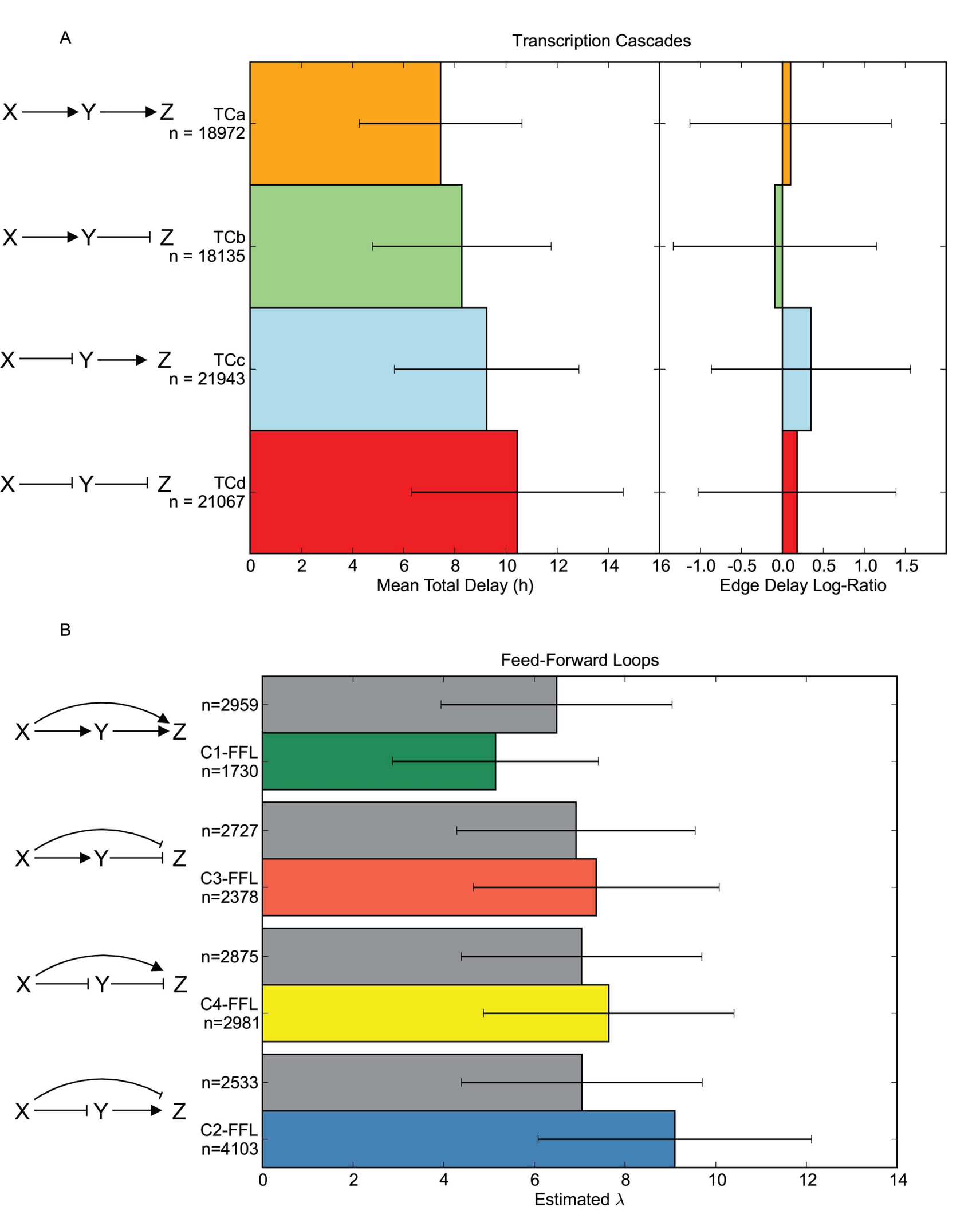
Slow Regulators

Flybase ID	Gene Symbol	λ	Random Mean	Random Std	z-score
FBgn0000099	ap	0.147	0.2257	0.0029	-27.04
FBgn0001077	ftz	0.177	0.2261	0.0037	-13.11
FBgn0004795	retn	0.167	0.2264	0.0048	-12.42
FBgn0003687	Tbp	0.127	0.2260	0.0087	-11.36
FBgn0259211	grh	0.134	0.2261	0.0119	-7.76
FBgn0000166	bcd	0.153	0.2272	0.0132	-5.61
FBgn0011656	Mef2	0.144	0.2289	0.0155	-5.49


Expression Correlation and ImaGO Co-Localization
Fraction
0.0
0.02
0.04
0.06
0.08
Corrected Predited Fraction in ImaGO
0.0
0.2
0.4
0.6
0.8
1.0
True Network Pearson
Random Network Pearson
True Network Step Function
Random Network Step Function
-0.8<= r <-0.7
-0.7<= r <-0.6
-0.6<= r <-0.5
-0.5<= r <-0.4
-0.4<= r <-0.3
-0.3<= r <-0.2
-0.2<= r <-0.1
-0.1<= r <0.0
0.0<= r <0.1
0.1<= r <0.2
0.2<= r <0.3
0.3<= r <0.4
0.4<= r <0.5
0.5<= r <0.6
0.6<= r <0.7
0.7<= r <0.8
0.8<= r <0.9
0.9<= r <1.0
Pearson Correlation


A
CG3225 (FBgn0031631)
Normalized Log Expression
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
sna
brk
sc
en
run
ase
ftz
retn
repo
Hr46
ap
kay
Dfd
5
10
15
20
Time (h)


B
Regulators and Target Mutual Information
True Regulators n = 5944
Randomized Regulators n = 300
Frequency of Target Profiles
0.3
0.2
0.1
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Mutual Information


A
Transcription Cascades
X → Y → Z
TCa
n = 18972
X → Y ⊣ Z
TCb
n = 18135
X ⊣ Y → Z
TCc
n = 21943
X ⊣ Y ⊣ Z
TCd
n = 21067
0 2 4 6 8 10 12 14
Mean Total Delay (h)
16 -1.0 -0.5 0.0 0.5 1.0 1.5
Edge Delay Log-Ratio
B
Feed-Forward Loops
n=2959
C1-FFL
n=1730
n=2727
C3-FFL
n=2378
n=2875
C4-FFL
n=2981
n=2533
C2-FFL
n=4103
0 2 4 6 8 10 12 14
Estimated λ

Supplemental Figure 1 – Fitted multi-step functions state change histogram. Setting S_{$_{max}$}=7 we found that none of the profiles had more than our state changes, attesting that a high enough threshold was chosen, about 45% of the expression profiles have at least 2 state changes meaning previous step function methods would fail to correctly discretize this dataset.

Supplemental Table 1 – GO term enrichment of all the fast regulators target genes using all annotated genes in the microarray as a background group.

Supplemental Table 2 – GO term enrichment of all the slow regulators target genes using all annotated genes in the microarray as a background group.

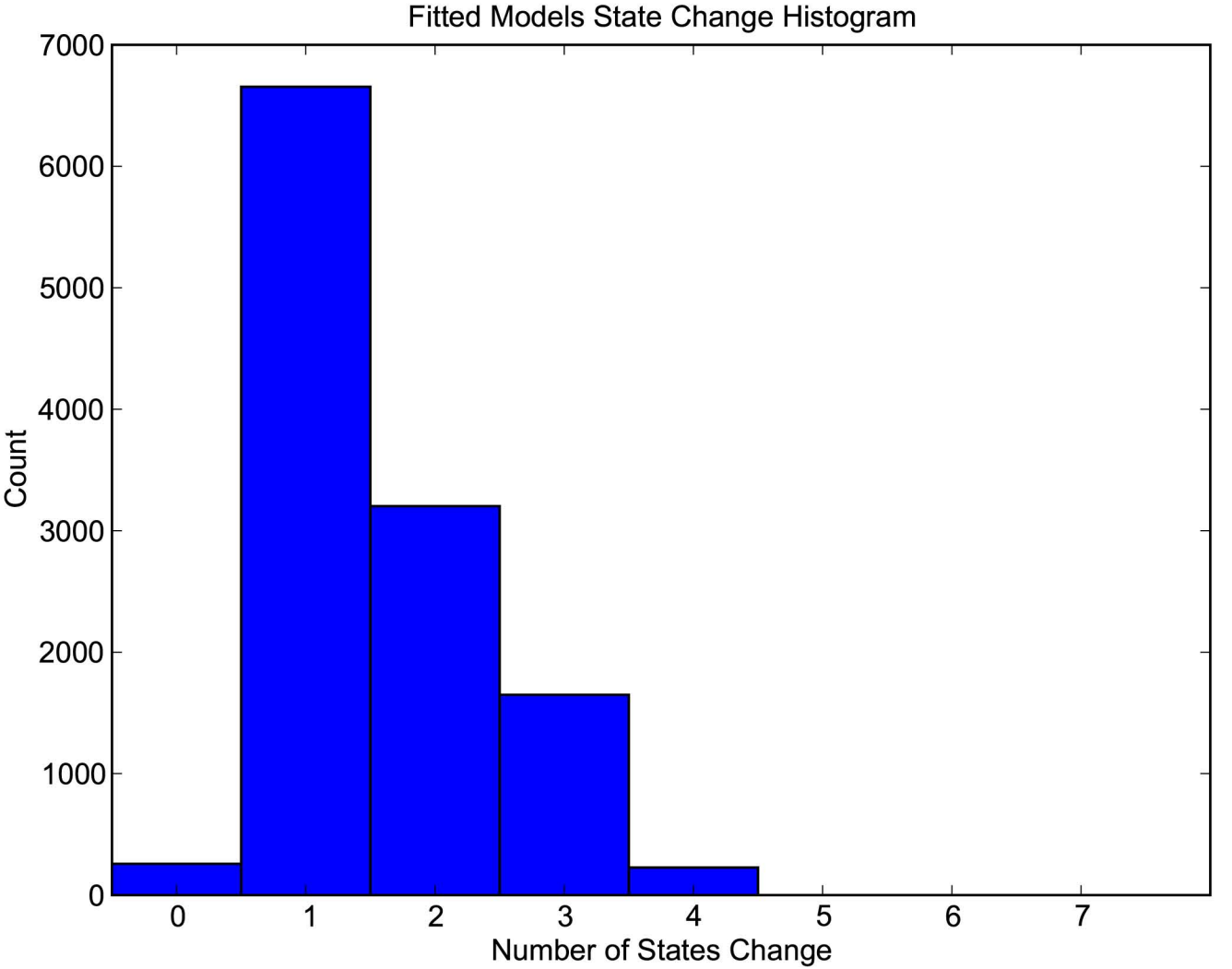
Supplemental Table 3 – Pearson correlation and multi-step functions predictions versus ImaGO contingency tables, for both the true regulatory network and for a random pairs of edges we present the contingency table of both the Pearson correlation and multi-step functions co-localization predictions against the ImaGO annotation, as well as several statistical measurements. Random pairs were pooled at about 5 times the number of true edges.

Supplemental Table 4 – Transcription cascade Kolmogorov-Smirnov test, all types of transcription cascade, a two-sample Kolmogorov-Smirnov test was performed to check if they have different delay distributions.

Supplemental Table 5 – Feed-forward loop Kolmogorov-Smirnov test, the four types of feed-forward loops with more than 50 indentified instances a two-sample Kolmogorov-Smirnov test verified that they do exhibit different λ distributions.

Supplemental Table 6 – Feed-forward loop Kolmogorov-Smirnov test against instances in a random network, the four types of feed-forward loops with more than 50 indentified instances a two-sample Kolmogorov-Smirnov test verified that they exhibit different distribution compared to random instances.

Supplemental Table 7 – Feed-forward loop target GO term enrichment. Enrichment was computed using Gorilla for a given set of genes using a background set of all the genes in the microarray. Only the top 15 most enriched GO terms for each motif type are shown.


Fitted Models State Change Histogram
Count
7000
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
Number of States Change

GO Term	Description	P-value
GO:0016192	vesicle-mediated transport	1.25E-06
GO:0007264	small GTPase mediated signal transduction	1.65E-05
GO:0006333	chromatin assembly or disassembly	3.57E-05
GO:0048193	Golgi vesicle transport	3.77E-05
GO:0051234	establishment of localization	9.20E-05
GO:0006810	transport	2.09E-04
GO:0045167	asymmetric protein localization during cell fate commitment	4.90E-04
GO:0045184	establishment of protein localization	5.48E-04

GO Term	Description	P-value
GO:0006333	chromatin assembly or disassembly	1.23E-10
GO:0009408	response to heat	7.08E-07
GO:0009266	response to temperature stimulus	1.02E-06
GO:0034605	cellular response to heat	1.75E-06
GO:0035080	heat shock-mediated polytene chromosome puffing	1.75E-06
GO:0035079	polytene chromosome puffing	1.75E-06
GO:0006325	establishment or maintenance of chromatin architecture	3.28E-06
GO:0006950	response to stress	1.14E-05
GO:0051276	chromosome organization	1.74E-05
GO:0050896	response to stimulus	1.09E-04
GO:0005975	carbohydrate metabolic process	5.48E-04

True Network Pearson Correlation		
	ImaGO Co-Localized	**ImaGO not Co-Localized**
Enhancer	1909	1923
Repressor	993	1871
Total	6696	
Sensitivity	0.657822	
Specificity	0.493147	
Precision	0.498173	
Recall	0.66655	
f-score	0.570191	
Accuracy	0.564516	

True Network Multi-Step Functions		
	ImaGO Co-Localized	**ImaGO not Co-Localized**
Enhancer	1717	1852
Repressor	1185	1942
Total	6696	
Sensitivity	0.591661	
Specificity	0.511861	
Precision	0.481087	
Recall	0.549089	
f-score	0.512843	
Accuracy	0.546446	

Random Pairs Pearson Correlation		
	ImaGO Co-Localized	**ImaGO not Co-Localized**
Enhancer	21058	38253
Repressor	13189	27267
Total	99767	
Sensitivity	0.614886	
Specificity	0.416163	
Precision	0.355044	
Recall	0.520516	
f-score	0.422144	
Accuracy	0.484379	

Random Pairs Multi-Step Functions		
	ImaGO Co-Localized	**ImaGO not Co-Localized**
Enhancer	13941	23197
Repressor	10869	21414
Total	69421	
Sensitivity	0.561911	
Specificity	0.480016	
Precision	0.375384	
Recall	0.431837	
f-score	0.401636	
Accuracy	0.509284	

	TCa		TCb		TCc		TCd	
	KS D	p-Value	KS D	p-Value	KS D	p-Value	KS D	p-Value
TCa n = 18972	0	1	0.105847	6.98E-91	0.097797	3.73E-85	0.042773	2.51E-16
TCb n = 18135			0	1	0.190345	1.186E-313	0.117048	1.11E-116
TCc n = 21943					0	1	0.11245	9.45E-119
TCd n = 21067							0	1

	C1-FFL		C2-FFL		C3-FFL		C4-FFL	
	KS D	p-Value	KS D	p-Value	KS D	p-Value	KS D	p-Value
C1-FFL n=1730	0	1	0.429128	1.90E-196	0.291981	3.67E-75	0.290301	3.72E-81
C2-FFL n=4103			0	1	0.252081	4.80E-84	0.202094	4.89E-62
C3-FFL n=2378					0	1	0.103028	1.05E-12
C4-FFL n=2981							0	1

	KS D	p-Value
C1-FFL n=1730	0.201539	3.14E-39
C2-FFL n=4103	0.238388	3.27E-78
C3-FFL n=2378	0.167843	1.01E-31
C4-FFL n=2981	0.132613	6.33E-23

A) C1-FFL

GO Term	Description	P-value
GO:0032502	developmental process	2.30E-20
GO:0048856	anatomical structure development	2.10E-19
GO:0050794	regulation of cellular process	2.00E-18
GO:0009653	anatomical structure morphogenesis	2.14E-18
GO:0050789	regulation of biological process	8.48E-17
GO:0048513	organ development	1.41E-16
GO:0065007	biological regulation	1.62E-15
GO:0045449	regulation of transcription	2.99E-13
GO:0010556	regulation of macromolecule biosynthetic process	7.22E-13
GO:0009889	regulation of biosynthetic process	7.28E-13
GO:0031326	regulation of cellular biosynthetic process	7.28E-13
GO:0031323	regulation of cellular metabolic process	2.38E-12
GO:0006355	regulation of transcription, DNA-dependent	2.51E-12
GO:0019219	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	3.05E-12
GO:0007389	pattern specification process	4.60E-12

B) C2-FFL

GO Term	Description	P-value
GO:0006030	chitin metabolic process	1.10E-05
GO:0007186	G-protein coupled receptor protein signaling pathway	4.02E-05
GO:0005976	polysaccharide metabolic process	4.30E-05
GO:0044264	cellular polysaccharide metabolic process	4.30E-05
GO:0006040	amino sugar metabolic process	5.52E-05
GO:0006041	glucosamine metabolic process	5.52E-05
GO:0006044	N-acetylglucosamine metabolic process	5.52E-05
GO:0006538	glutamate catabolic process	8.73E-04
GO:0007155	cell adhesion	8.89E-04

C) C3-FFL

GO Term	Description	P-value
GO:0001709	cell fate determination	4.01E-06
GO:0007600	sensory perception	5.49E-06
GO:0001700	embryonic development via the syncytial blastoderm	6.23E-06
GO:0009792	embryonic development ending in birth or egg hatching	6.65E-06
GO:0048856	anatomical structure development	1.12E-05
GO:0007606	sensory perception of chemical stimulus	2.41E-05
GO:0009653	anatomical structure morphogenesis	3.43E-05
GO:0048513	organ development	4.39E-05
GO:0006355	regulation of transcription, DNA-dependent	6.88E-05
GO:0051252	regulation of RNA metabolic process	9.37E-05
GO:0007419	ventral cord development	1.34E-04
GO:0032502	developmental process	2.05E-04
GO:0035287	head segmentation	2.65E-04
GO:0035288	anterior head segmentation	3.45E-04
GO:0050877	neurological system process	3.75E-04

C) C4-FFL

GO Term	Description	P-value
GO:0005976	polysaccharide metabolic process	2.88E-06
GO:0044264	cellular polysaccharide metabolic process	2.88E-06
GO:0006030	chitin metabolic process	3.42E-06
GO:0006040	amino sugar metabolic process	5.02E-06
GO:0006041	glucosamine metabolic process	5.02E-06
GO:0006044	N-acetylglucosamine metabolic process	5.02E-06
GO:0007186	G-protein coupled receptor protein signaling pathway	3.19E-05
GO:0007218	neuropeptide signaling pathway	2.13E-04
GO:0019236	response to pheromone	3.66E-04
GO:0032501	multicellular organismal process	4.89E-04
GO:0042221	response to chemical stimulus	7.38E-04
GO:0050896	response to stimulus	7.44E-04