DREISS: Using state-space models to infer the dynamics of gene expression driven by external and internal regulatory networks

Daifeng Wang[1,2], Fei He[4], Sergei Maslov[4], Mark Gerstein[1,2,3]*

[1]Program in Computational Biology and Bioinformatics; [2]Department of Molecular Biophysics and Biochemistry; [3]Department of Computer Science, Yale University, New Haven, CT, USA. [4]Biological, Environmental and Climate Sciences Department, Brookhaven National Laboratory, Upton, NY, USA.

*Correspondence to: mark@gersteinlab.org

ABSTRACT

Gene expression is controlled by the combinatorial effects of regulatory factors from different biological subsystems such as general transcription factors, cellular growth factors and microRNAs. A subsystem's gene expression may be controlled by its internal regulatory factors, exclusively, or by external subsystems, or by both. It is thus useful to distinguish the degree to which a subsystem is regulated internally or externally – e.g., how species-specific regulatory factors affect the expression of conserved genes during evolution.

We developed a computational method (DREISS) for dynamics of gene expression driven by external and internal regulatory programs based on state space models to help dissect the effects of different regulatory subsystems on gene expression (dreiss.gersteinlab.org). Given a subsystem, the "state" and "control" in the model refer to its own (internal) and another subsystem's (external) gene expression levels. The state at a given time is determined by the state and control at a previous time. Because typical time-series data do not have enough samples to estimate the model's parameters, DREISS uses dimensionality reduction, and identifies canonical temporal expression trajectories (e.g., degradation, growth and oscillation) representing the regulatory effects coming from various subsystems.

To demonstrate capabilities of DREISS, we study the regulatory effects of evolutionarily conserved vs. divergent transcription factors across distant species. In particular, we applied DREISS to the time-series gene expression datasets of *C. elegans* and *D. melanogaster* during their embryonic development. We analyzed the expression dynamics of the conserved, orthologous genes (orthologs), seeing the degree to which these can be accounted for by orthologous (internal) versus species-specific (external) transcription factors (TFs). We found that between two species, the orthologs have matched internally driven dynamic expression patterns but very different externally driven patterns. This is particu-

larly true for genes with evolutionarily ancient functions (e.g. the ribosomal proteins), in contrast to those with more recently evolved functions (e.g., cell-cell communication). This suggests that despite striking morphological differences, some fundamental embryonic-developmental processes are still controlled by ancient regulatory systems.

AUTHOR SUMMARY

The dynamics of a biological system can be controlled by its own internal mechanisms and external perturbations. To gain intuition on this, we may draw a comparison with a mass hanging from a spring. The mass will move naturally by itself but its dynamics is also affected by one's pulling it. That is, the dynamics of the mass is governed by the effect of the external perturbations superimposed on the internal mechanism of the spring (i.e. Hooke's law). Similarly, given a group of genes, their temporal gene expression dynamics can be controlled by both transcription factors inside the group and external regulatory factors. Therefore, it is useful to identify the expression dynamics that are exclusively controlled by internal or external factors and compare them across various systems. While state-space models have been widely used to decouple the internal and external effects in physical systems, such as the mass and spring, typical biological systems do not have enough time samples to infer all the model's parameters, and applications of state-space models have not very effective in these instances. Hence, we developed a general-purpose computational method by integrating state-space models and dimensionality reduction to identify temporal gene expression patterns driven by internal and external regulatory networks. We applied our method to the embryonic developmental datasets in the worm and fly (and also in a human cancer context). We successfully identified the temporal expression dynamics of cross-species conserved genes that were driven by conserved and species-specific regulatory networks.

1 INTRODUCTION

Gene regulatory networks systematically control the gene expression dynamics. These networks are highly modular, and consist of various sub-networks. Each sub-network contains a number of regulatory factors representing a subsystem that drives specific gene regulatory functions [1,2]. The subsystems interact with one another, and work together to carry out the entire gene regulatory function. For example, the gene expression in embryogenesis is controlled by the combinatorial effects of various regulatory subsystems composed of complex evolutionary regulatory networks [3]. These regulatory subsystems

drive very diverse developmental programs, from the highly conserved (e.g. DNA replication) to the species-specific (e.g. body segmentation). As such the orthologous genes that are evolutionary conserved genes across species can therefore be regulated by both orthologous and species-specific transcription factors (TFs) [4]. The orthologous TFs form an "internal" regulatory network, while the species-specific TFs form an "external" one. Unfortunately, existing experimental gene expression data cannot decouple the expression components that are driven by the different subsystems. Thus, computational methods are required to assess the contribution from each factor or subsystem from the gene expression data. In this study, we propose a novel computational method, DREISS - dynamics of gene expression driven by external and internal regulatory networks based on state space model. Using DREISS, we are able to identify temporal gene expression dynamic patterns for evolutionarily conserved genes during embryonic development, as driven by conserved and species-specific regulatory subsystems. These results advance our current understanding of gene regulatory networks during evolution, as well as the differentiation during development.

Developmental gene regulatory networks control gene expression during the developmental processes. These particular regulatory networks have evolved, making it difficult to understand their regulatory mechanisms at the system level. Hence, one typically compares developmental gene expression across species to infer biological activities of developmental gene regulatory networks. For example, embryogenesis provides a platform to study the evolution of gene expression between different species. Recent work has showed that significant biological insight can be gained by cross-species comparisons of the expression profiles during embryogenesis for worms [5], flies [6], frogs [7] and several other vertebrates [8]. It was found that the orthologous genes have minimal temporal expression divergence during the phylotypic stage, a middle phase during the embryonic development across species within the same phylum. These patterns are often characterized as "hourglass" [9]. In addition, the conserved hourglass patterns were observed even within a single species while comparing the developmental gene expression data across distant species, such as worm and fly [10]; i.e., the expression divergence among evolutionarily conserved genes become minimal during the phylotypic stage in both worm and fly. However, much less is known about how the orthologous genes in each species eventually contribute to their species-specific phenotypes due to the lack of appropriate computational approaches. Thus, we aim to use DREISS to discover the components of the orthologous gene expression during embryonic development driven by species-specific transcription factors.

The state-space model has been widely used in engineering [11], and also in biology for the analysis of gene expression dynamics [12-14]. It models the dynamical system output as a function of both the current internal system state and the external input signal. A commonly used example in engineering is the vehicle cruise control system where the system state can be the vehicle's speed. Based on the road conditions, the cruise control requires various fuel amounts in order to keep the desired speed level. In biology, we can look at the transcription factors and microRNAs as internal and respectively external regulatory factors of the protein-coding genes expression (See more internal-external examples in Supplemental Table 1). Similarly, the state-space model can be applied for studying the expression of orthologous genes at different developmental stages using information regarding their expression (internal) and species-specific regulatory factors (external) at the current known developmental stage. Unlike earlier studies that calculate the expression correlation between individual genes, the state-space model predicts the temporal causal relationships at the system level; i.e., the state at a time is determined by the state and external input at the previous time. The earlier work applied the state-space model to study the gene expression dynamics focusing on small-scale systems, and did not explore the analytic dynamic characteristics of the inferred state-space models. The complex and large-scale biological datasets, especially temporal gene expression data, are very noisy, and high dimensional (i.e., the number of genes is much greater than the number of time samples), thereby preventing an accurate estimation of the state-space model's parameters. The dimensionality reduction techniques have thus been used to project high-dimensional genes to low-dimensional meta-genes (i.e., the selected features representing de-noised and systematic expression patterns [1,15,16]) as well as the principal dynamic patterns for those meta-genes [17,18]. Using DREISS, we are able to apply the dimensionality reduction to the gene expression data, and develop an effective state-space model for their meta-genes, and then identify a group of canonical temporal expression trajectories representing the dynamic patterns driven by the effective conserved and species-specific meta-gene regulatory networks according to the model's analytic characteristics. These dynamic patterns reveal temporal gene expression components that are controlled by conserved or species-specific GRNs.

DREISS is a general-purpose tool and can be used to study the gene regulatory effects from any different subsystems for a given group of genes. As an illustration, we applied DREISS to the gene expression data during embryonic development for two model organisms, worm (*Caenorhabditis elegans*) and fly

Daifeng Wang 3/29/16 3:23 PM
Deleted: output and

Daifeng Wang 3/29/16 3:23 PM
Deleted: is

(*Drosophila melanogaster*). In both species, we were able to identify the expression patterns of worm-fly orthologs driven by the conserved regulatory network consisting of the worm-fly orthologous TFs (i.e., the conserved regulatory subsystems between two species), as well as the worm/fly-specific regulatory network consisting of non-orthologous TFs (i.e., the species-specific regulatory subsystem). Our results reveal that, in addition to executing conserved developmental functions between worm and fly, their orthologous genes are also regulated by species-specific TFs to involve in species-specific developmental processes. In summary, DRIESS provides a framework to analyze both distantly and closely related species allowing for a better understanding of the gene regulatory mechanisms during development.

2 METHODS

DREISS consists of five major steps as detailed in Figure 1:

Step A: DREISS models temporal gene expression dynamics using state-space models in control theory. In this step, we need to define the internal and external groups of genes and input their time-series gene expression data that we are interested to study. We assume that the time-series gene expression data fits a state-space module. In the state-space model, the "state" refers to the expressions for a large group of genes of interest, such as the worm-fly orthologous genes investigated here. The "control" refers to any other group of genes that contribute to the gene expression of the "state", such as the species-specific TFs contributed to control orthologous gene expression.

Step B: Due to the limited number of temporal samples in gene expression experiments, we do not have enough data to accurately estimate the parameters of the state-space models that capture interactions among hundreds of genes. Therefore, DREISS projects high-dimensional gene expression space to lower-dimensional meta-gene expression spaces using dimensionality reduction techniques.

Step C: DREISS derives the effective state-space models for meta-genes so that model parameters can be estimated.

Step D: DREISS identifies the meta-gene expression dynamic patterns; i.e., canonical temporal expression trajectories driven by "state" (internal) and by "control" (external) based on the analytic solutions of the estimated models.

Step E: Finally, DREISS calculates the gene coefficients over canonical temporal expression trajectories based on linear transformations between genes and meta-genes. DREISS also allows us to compare the

dynamic expression patterns of multiple datasets with samples taken at different times. We describe each DREISS step in detail as follows.

2.1 State-space models for temporal gene expression dynamics

A gene regulatory network is made up of various subsystems [1,2]. These subsystems work together to execute regulatory functions. Given a group of N_1 genes in a subsystem, defined as the internal gene set, Ω, their gene expression levels are not only controlled by internal interactions among Ω, but also affected by the regulatory factors from other subsystems outside Ω. We define an external gene set, Ψ consisting of those external regulatory factors. For example, we consider the worm-fly orthologous genes as internal set Ω. The worm-fly orthologous TFs from internal set Ω are the *internal* regulatory factors, and non-orthologous TFs such as worm- or fly- specific TFs are the *external* regulatory factors. Both the internal and external regulatory factors control gene expressions in dynamic ways (i.e., their regulatory signals at the current time will affect gene expressions at subsequent times). Thus, the regulatory mechanisms for the gene expressions form a control system. In this study, we used a state-space model (defined by linear first-order difference equations, Figure 2A) to formulate temporal gene expression dynamics for internal set Ω (comprising N_1 genes) with external regulation parameters from external set Ψ (comprising N_2 genes) at time points *1*, *2*, … , *T* as follows:

$$X_{t+1} = AX_t + BU_t \quad (1)$$

, where the vector $X_t \in \Re^{N_1 \times 1}$, the "state", includes N_1 gene expression levels at time t in Ω, and the vector $U_t \in \Re^{N_2 \times 1}$, the "input or control", includes N_2 gene expression levels at time t in Ψ. The system matrix $A \in \Re^{N_1 \times N_1}$ captures internal causal interactions among genes in Ω (i.e., the i^{th}, j^{th} element of A, A_{ij} describes the contribution from the j^{th} gene expression at time t to the i^{th} gene expression at the next time t+1), which instantiates a gene regulatory network. The control matrix $B \in \Re^{N_1 \times N_2}$ captures external causal regulations from the genes in Ψ to genes in Ω (i.e., the i^{th}, j^{th} element of B, B_{ij} describes the contribution from the j^{th} gene expression in Ψ at time t to the i^{th} gene expression in Ω at the next time t+1). $\Re$ represents the real number domain. According to the state space model (1), the gene expression dynamics in Ω is determined by the system matrix A and the control matrix B.

2.2 Dimensionality reduction from genes to meta-genes

The temporal gene expression experiments normally have limited time samples (for example, there may only be a dozen time points), which are far less than the time samples needed to estimate the large ma-

trices A and B when internal and external groups, Ω and Ψ are composed of hundreds or thousands of genes. One way to deal with lack of time samples is dimensionality reduction. Thus, we project high dimensional temporal gene expressions to much lower dimensional meta-gene expression levels using a dimensionality reduction technique (Figure 2B). Those meta-gene expression levels should capture original gene expression patterns, such as the ones having the greatest degree of co-variation. We calculate the meta-gene expression levels as follows:

$\tilde{X}_t = W_X^* X_t; \tilde{U}_t = W_U^* U_t$ (2)

, where $\tilde{X}_t \in \Re^{M_1 \times 1}$, the "meta-gene state" at time t, includes M_1 ($<< N_1$ and $<T$) meta-gene expression levels; i.e., the first M_1 elements of the t^{th} row of the matrix whose columns are right-singular vectors of the matrix $[X_1 X_2 ... X_T]$ in Ω by the singular value decomposition (SVD) [19]; the vector $\tilde{U}_t \in \Re^{M_2 \times 1}$, the "meta-gene input or control" at time t, includes M_2 ($<< N_2$ and $<T$) meta-gene expression levels; i.e., the first M_2 elements of the t^{th} row of the matrix whose columns are right-singular vectors from SVD of the matrix $[U_1 U_2 ... U_T]$ in Ψ; $W_X \in \Re^{N_1 \times M_1}$ is the linear projection matrix of SVD from M_1 meta-gene expression space to N_1 gene expression space in Ω, $W_U \in \Re^{N_2 \times M_2}$ is the linear projection matrix of SVD from M_2 meta-gene expression space to N_2 gene expression space in Ψ, and $(.)^*$ is a pseudo-inverse operation; i.e., $W^*W=I$, where I is the identity matrix.

2.3 Estimation of effective state-space model for meta-gene expression dynamics

Next, we obtain the effective state-space model for meta-genes using linear projections W_X and W_U between genes and meta-genes as follows (Figure 2C). By replacing (1) using (2), we obtain that

$W_X \tilde{X}_{t+1} = A W_X \tilde{X}_t + B W_U \tilde{U}_t$. (3)

After multiplying the pseudo-inverse of W_X, $W_X^* \in \Re^{M_1 \times N_1}$ s.t. $W_X^* W_X = I$ where I is an identity matrix, at both sides of (3), we have that

$$\tilde{X}_{t+1} = \underbrace{W_X^* A W_X}_{\tilde{A}} \tilde{X}_t + \underbrace{W_X^* B W_U}_{\tilde{B}} \tilde{U}_t \Rightarrow \tilde{X}_{t+1} = \tilde{A}\tilde{X}_t + \tilde{B}\tilde{U}_t \quad (4)$$

, where the effective meta-gene system matrix $\tilde{A} = W_X^* A W_X \in \Re^{M_1 \times M_1}$ captures internal causal interactions among meta-genes in Ω (i.e., an element of $\tilde{A}$, $\tilde{A}_{ij}$ describes the contribution from the j^{th} meta-gene expression at time t to i^{th} meta-gene expression at time t+1), and the effective control matrix $\tilde{B} = W_X^* B W_U \in \Re^{M_1 \times M_2}$ captures external causal regulations from meta-genes of Ψ to meta-genes of Ω

(i.e., the i^{th}, j^{th} element of $\tilde{B}$, $\tilde{B}_{ij}$ describes the contribution from the j^{th} meta-gene expression in Ψ at time t to i^{th} meta-gene expression in Ω at time t+1). Equation (4) describes the effective state space model for the meta-genes of Ω, whose expression dynamics is determined by $\tilde{A}$ and $\tilde{B}$. Because the meta-gene dimension, M_1 (M_2) is less than T, and much less than N_1 (N_2), we can estimate $\tilde{A}$ and $\tilde{B}$ as follows.

We rewrite Equation (4) as a matrix product on the right side:

$$\tilde{X}_{t+1} = \tilde{A}\tilde{X}_t + \tilde{B}\tilde{U}_t = \begin{bmatrix} \tilde{A} & \tilde{B} \end{bmatrix} \begin{bmatrix} \tilde{X}_t \\ \tilde{U}_t \end{bmatrix}. \qquad (5)$$

By applying Equation (5) to time points, 2,3, … , T, we then obtain that

$$\underbrace{\begin{bmatrix} \tilde{X}_2 & \tilde{X}_3 & \cdots & \tilde{X}_T \end{bmatrix}}_{\mathrm{Z}} = \begin{bmatrix} \tilde{A} & \tilde{B} \end{bmatrix} \underbrace{\begin{bmatrix} \tilde{X}_1 & \tilde{X}_2 & \cdots & \tilde{X}_{T-1} \\ \tilde{U}_1 & \tilde{U}_2 & \cdots & \tilde{U}_{T-1} \end{bmatrix}}_{\Upsilon} \qquad (6)$$

, where $\mathrm{Z} \in \Re^{M_1 \times (T-1)}$ and $\Upsilon \in \Re^{(M_1+M_2)\times(T-1)}$.

Because of dimension reduction, Υ has more columns than rows so that it has right pseudo-inverse. Thus, the effective internal system matrix $\tilde{A}$ and external control matrix $\tilde{B}$ can be estimated by:

$$\begin{bmatrix} \tilde{A} & \tilde{B} \end{bmatrix} = \mathrm{Z}\Upsilon^{*} \qquad (7)$$

, where $\Upsilon^{*} \in \Re^{(T-1)\times(M_1+M_2)}$ is the right pseudo-inverse of Υ; i.e. $\Upsilon\Upsilon^{*} = I$, with $M_1 < N_1, M_2 < N_2, M_1 + M_2 < T, t = 1,2,...,T$. It is worth noting that if we do not reduce the dimensionality, and obtain Equation (6) from Equation (4), then Υ will have much more rows than columns so that it doesn't have right pseudo-inverse; i.e., there doesn't exist a matrix Υ^{*} such that $\Upsilon\Upsilon^{*}$ is a full-rank identify matrix. In addition, the condition of $M_1+M_2<T$ also makes $\Upsilon\Upsilon^{*}$ is a full-rank identify matrix.

Daifeng Wang 3/29/16 3:23 PM
Deleted: Υ

2.4 Identification of internally and externally driven principal dynamic expression patterns of meta-genes (canonical temporal expression trajectories)

The analytic solution to a general first-order linear matrix difference equation [20], $Q_{t+1}=CQ_t$ is

$Q_t=C^tQ_0=(HEH^{-1})^tQ_0=HE^tH^{-1}Q_0=HE^tS$, where the columns of the matrix H are eigenvectors of C, the diagonal elements of the diagonal matrix E are eigenvalues of C such that $CH=HE$, and the vector $S= H^{-1}Q_0$. Then, if we rewrite Q_t by a linear combination of the time exponential of eigenvalues of C, we have that $Q_t = HE^tS = \sum_{i=1}^{m_c} \alpha_i^t s_i H_i = \sum_{i=1}^{m_c} \alpha_i^t K_i$, where m_c is the total number of eigenvalues of C, α_i is the i^{th} eigenvalue of C, s_i is the i^{th} element of S, H_i is the i^{th} eigenvector of C (i.e., the i^{th} column of H), and $K_i=s_iH_i$ is the coefficient vector of Q_t over the t^{th} time exponential of α_i.

From Equation (4), the internally driven components of meta-gene states at two adjacent time points have $\tilde{X}_{t+1}^{\text{INT}} = \tilde{A}\tilde{X}_t^{\text{INT}} \in \Re^{M_1\times 1}$. According to the above analytic solution, the components of meta-gene expressions in Ω driven by effective internal regulations are linear combinations of M_1 dynamic patterns determined by the eigenvalues of the effective system matrix $\tilde{A}$ as follows:

$\tilde{X}_t^{\text{INT}} = \sum_{p=1}^{M_1} \lambda_p^t \tilde{K}_p$; i.e., the internally driven component of i^{th} meta-gene's expression across all time points, $[\tilde{X}_1^{\text{INT}}(i) \quad \tilde{X}_2^{\text{INT}}(i) \quad \dots \quad \tilde{X}_T^{\text{INT}}(i)] = \sum_{p=1}^{M_1} \tilde{K}_p(i) \underbrace{[\lambda_p^1 \quad \lambda_p^2 \quad \dots \quad \lambda_p^T]}_{p^{\text{th}}\ \text{iPDP}}$ (8)

, where λ_p and $\tilde{K}_p \in \mathbb{C}^{M_1\times 1}$ are the p^{th} eigenvalue of $\tilde{A}$ and its coefficient vector from the analytic solution, which determines the p^{th} dynamic pattern driven by effective internal regulations, defined as the p^{th} internal principal dynamic pattern (iPDP) = $[\lambda_p^1\ \lambda_p^2\ \dots\ \lambda_p^T]$, in which λ_p^t represents the t^{th} power of λ_p, and $\Xi(i)$ represents i^{th} element of the vector Ξ. $\mathbb{C}$ represents the complex number domain. If an eigenvalue λ is complex when $\tilde{A}$ is asymmetric, then its conjugate $\bar{\lambda}$ is also an eigenvalue, so we sum its iPDP and its conjugate eigenvalue, $\bar{\lambda}$'s iPDP, as a unified iPDP with real elements equal to $[\lambda_p^1+\bar{\lambda}_p^1\ \lambda_p^2+\bar{\lambda}_p^2\ \dots\ \lambda_p^T+\bar{\lambda}_p^T]$.

The internal principal dynamic patterns (iPDPs) represent canonical temporal expression trajectories, which can be either increasing, or damped oscillation and so on depending on jPDP's eigenvalues (Table 1). The jPDPs can be ordered by sorting their eigenvalues.

Table 1. Classification of canonical temporal expression trajectories for jPDP eigenvalue types

PDP eigenvalue	Real						Complex (radius)		
	>1	=1	<1 & >0	<0 & > -1	= -1	<-1	>1	=1	<1
Canonical temporal expression	Increasing (I)	Flat (F)	Decreasing (D)	Vibrating early (VE)	Vibrating (V)	Vibrating late (VL)	Underdamped oscillation	Oscillation (O)	Damped oscillation (DO)

Daifeng Wang 3/29/16 3:23 PM
Moved down [1]: components of meta-gene expressions in Ω driven by effective external regulations from Ψ, i.e., $\tilde{X}_{t+1}^{\text{EXT}} = \tilde{B}$

Daifeng Wang 3/29/16 3:23 PM
Deleted: Similarly, the

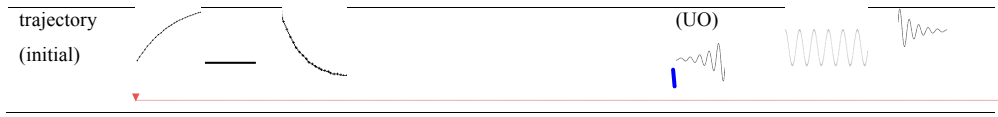
Daifeng Wang 3/29/16 3:23 PM
Deleted: $\tilde{X}_t^{\text{EXT}} \in \Re^{M_2\times 1}$ (externally driven components of meta-gene states at two adjacent time points) are linear combinations of M_2 dynamic patterns determined by the eigenvalues of the effective system matrix $\tilde{B}$ as follows:

Daifeng Wang 3/29/16 3:23 PM
Deleted: PDP's

Daifeng Wang 3/29/16 3:23 PM
Deleted: PDPs

Daifeng Wang 3/29/16 3:23 PM
Deleted:

Daifeng Wang 3/29/16 3:23 PM
Deleted: PDP



The components of meta-gene expressions in Ω driven by effective external regulations from Ψ, i.e., $\tilde{X}_{t+1}^{\mathrm{EXT}} = \tilde{B}\tilde{U}_t \in \Re^{M_1 \times 1}$ (externally driven components of meta-gene states at two adjacent time points) are as follows: the externally driven component of i^{th} internal meta-gene's expression across time points,

$$[\tilde{X}_2^{\mathrm{EXT}}(i) \quad \tilde{X}_3^{\mathrm{EXT}}(i) \quad \dots \quad \tilde{X}_T^{\mathrm{EXT}}(i)] = \sum_{q=1}^{M_2} \tilde{B}_{i,q} \underbrace{[\tilde{U}_1(q) \quad \tilde{U}_2(q) \quad \dots \quad \tilde{U}_{T-1}(q)]}_{q^{\mathrm{th}}\ \mathrm{ePDP}} \qquad (9)$$

, where $\tilde{X}_t^{\mathrm{EXT}}(i)$ and $\tilde{U}_t(q)$ are i^{th} and q^{th} elements of $\tilde{X}_t^{\mathrm{EXT}}$ and $\tilde{U}_t$, respectively with t=1,2,…, T, the vector $\underbrace{[\tilde{U}_1(q) \quad \tilde{U}_2(q) \quad \dots \quad \tilde{U}_{T-1}(q)]}$ is defined as q^{th} external principal dynamic pattern (ePDP), and $\tilde{B}_{i,q}$ is the element of $\tilde{B}$ at i^{th} row and q^{th} column, which is also the coefficient of the externally driven component of i^{th} internal meta-gene's expression over q^{th} ePDP.

2.5 Identification of gene coefficients of principal expression dynamic patterns

Because genes and meta-genes have linear relationships in terms of their expression levels as described in Equation (2), the components of gene expression levels in Ω driven by internal regulations, $X_t^{\mathrm{INT}} \in \Re^{N_1 \times 1}$ can be also expressed as linear combinations of M_1 iPDPs:

$X_t^{\mathrm{INT}} = W_X \tilde{X}_t^{\mathrm{INT}} = \sum_{p=1}^{M_1} \lambda_p^t \underbrace{W_X \tilde{K}_p}_{C_p} = \sum_{p=1}^{M_1} \lambda_p^t C_p$; i.e.,

the internally driven component of i^{th} gene's expression across all time points,

$$[X_1^{\mathrm{INT}}(i) \quad X_2^{\mathrm{INT}}(i) \quad \dots \quad X_T^{\mathrm{INT}}(i)] = \sum_{p=1}^{M_1} C_p(i) \underbrace{[\lambda_p^1 \quad \lambda_p^2 \quad \dots \quad \lambda_p^T]}_{p^{\mathrm{th}}\ \mathrm{iPDP}} \qquad (10)$$

, where $C_p = W_X \tilde{K}_p \in \mathbb{C}^{M_1 \times 1}$ represents the gene coefficient vector for p^{th} iPDP. Similarly, the gene expression components driven by external regulations from Ψ, $X_t^{\mathrm{EXT}} \in \Re^{N_1 \times 1}$ can be also expressed as linear combinations of M_2 ePDPs:

$X_t^{\mathrm{EXT}} = W_X \tilde{X}_t^{\mathrm{EXT}} = \underbrace{W_X \tilde{B}}_{D} \tilde{U}_t = D\tilde{U}_t$; i.e.,

the externally driven component of i^{th} gene's expression across all time points,

$$[X_2^{\mathrm{EXT}}(i) \quad X_3^{\mathrm{EXT}}(i) \quad \dots \quad X_T^{\mathrm{EXT}}(i)] = \sum_{q=1}^{M_2} D_{i,q} \underbrace{[\tilde{U}_1(q) \quad \tilde{U}_2(q) \quad \dots \quad \tilde{U}_{T-1}(q)]}_{q^{\mathrm{th}}\ \mathrm{ePDP}} \qquad (11)$$

Daifeng Wang 3/29/16 3:23 PM
Deleted: <sp>

Daifeng Wang 3/29/16 3:23 PM
Moved (insertion) [1]

Daifeng Wang 3/29/16 3:23 PM
Deleted: $\sum_{q=1}^{M_2} \sigma_q^t \underbrace{W_X \tilde{L}_q}_{D_q} = \sum_{q=1}^{M_2} \sigma_q^t D_q$

Daifeng Wang 3/29/16 3:23 PM
Deleted: $[X_1^{\mathrm{EXT}}(i) \quad X_2^{\mathrm{EXT}}(i) \quad \dots \quad X_T^{\mathrm{EXT}}(i)]$

Daifeng Wang 3/29/16 3:23 PM
Deleted: $\sum_{q=1}^{M_2} D_q(i) \underbrace{[\sigma_q^1 \quad \sigma_q^2 \quad \dots \quad \sigma_q^T]}_{q^{\mathrm{th}}\ \mathrm{ePDP}}$

, where $X_t^{\mathrm{EXT}}(i)$ is i^{th} element of X_t^{EXT} with t=1,2,…, T, and $D_{i,q}$ is the element of $D = W_X\tilde{B}$ at i^{th} row and q^{th} column, which is also the coefficient of the externally driven component of i^{th} gene's expression over q^{th} ePDP.

Daifeng Wang 3/29/16 3:23 PM
Deleted: $D_q = W_X\tilde{L}_q \in \mathbb{C}^{M_2\times 1}$ represents

Daifeng Wang 3/29/16 3:23 PM
Deleted: gene

Daifeng Wang 3/29/16 3:23 PM
Deleted: vector for

3 RESULTS

Gene expression data during embryogenesis provide important information about the dynamics of genomic functions throughout the developmental process, from the conserved functions such as DNA replication to the species-specific functions such as body segmentation, but hardly reveal any data regarding the evolutionary gene regulatory subsystems that drive those developmental functions [3]. Thus, in order to understand the relationships between those subsystems and their driving genomic functions, we apply DREISS to worm and fly gene expression datasets during embryogenesis in modENCODE and we are able to identify various developmental genomic functions of worm-fly orthologous gene pairs driven by two different evolutionary regulatory subsystems, conserved (worm-fly orthologous TFs) and non-conserved (worm/fly specific TFs). As model organisms for developmental biology, both worm and fly have been used previously to study embryogenesis.

3.1 Applications to worm and fly embryonic developmental data in modENCODE: orthologous genes, transcription factors and gene expression datasets

DREISS enables us to compare expression dynamic patterns between two or more temporal gene expression datasets even though they have different numbers of samples, as well as differences in the times at which those samples were collected. For example, we can apply DREISS to two different datasets of the same group of genes, and identify both the common (similar) and the specific (different) dynamic patterns driven by internal regulations captured by the eigenvalues of the effective system matrices between the two datasets.

In this paper, we apply DREISS to 3,153 one-to-one orthologous genes between worm (*Caenorhabditis elegans*) and fly (*Drosophila melanogaster*) as internal group, Ω to study their expression dynamics during embryonic development [10]. We refer to species-specific TFs as external regulations; i.e., external group Ψ. We found that worm-fly orthologs have similar internal dynamic patterns, which may be mainly driven by conserved TFs, but have very different external dynamic patterns driven by species-specific TFs between worm and fly embryonic developmental stages. The data is summarized as follows.

We define internal group Ω as 3,153 one-to-one orthologous genes between worm and fly during embryonic development, and external group Ψ as all the species-specific TFs (509 worm-specific TFs, 442 fly-specific TFs) [21,22]. We used their temporal gene expression levels (as measured by the RPKM values in RNA-seq) during embryonic development from the modENCODE project [10]. The worm embryonic development dataset includes T=25 time stages at 0, 0.5, 1, 1.5, … , 12 hours, and the fly dataset includes T=12 time stages at 0, 2, 4, … , 22 hours, but t=1,2,..,25 for worm and t=1,2,…,12 for fly are used in this paper, representing the relative time points for the entire embryonic development processes. Because $M_1+M_2<T$ in Equation (7), we choose $M_1=M_2=5$ meta-genes for fly (T=12), and find that five meta-genes of Ω and five meta-genes of Ψ capture ~98% of the co-variation of orthologous gene expressions and fly-specific TF gene expressions, respectively. In order to compare worm and fly, we also choose $M_1=M_2=5$ meta-genes for worm, which capture ~98% of the co-variation of orthologous gene expressions and worm-specific TF gene expressions.

3.2 Meta-genes of worm-fly orthologous genes have similar internal, yet different external principal dynamic patterns during embryonic development

We find that the meta-gene canonical temporal expression trajectories driven by conserved regulatory networks (i.e., internal principal dynamic patterns, iPDPs) include four major patterns in both the worm and fly embryonic developmental process by order of eigenvalues: 1) a late highly varied pattern; 2) an early fast decaying pattern; 3) a slowly increasing pattern; and 4) an oscillating pattern (Figure 3A); i.e., the pattern of canonical trajectories (VL, D, I, O) in Table 1. In contrast to the observed iPDP similarities, we find that worm and fly have very different external principal dynamic patterns (ePDPs) (Figure 3B); i.e., the expression dynamic patterns driven by species-specific TFs. The patterns driven by the worm-specific regulatory network; i.e., worm ePDPs, include a varied pattern that decreases until the middle stage and then increases , an increasing pattern, a varied pattern with a peak entering middle stage, a pattern that varies early and then increases during the embryonic development, and a cosine-like oscillating pattern with roughly two periods during the embryonic development. The fly ePDPs, however, have a varied pattern with low expression at the early stage, a sine-like oscillating pattern with roughly one period during the embryonic development, an increasing pattern, another sine-lie oscillating pattern with roughly two periods during the embryonic development, and a varied pattern that is like damped oscillation. In addition, we checked the sensitivity of iPDPs to small perturbations to inter-

Daifeng Wang 3/29/16 3:23 PM
Deleted: canonical temporal

Daifeng Wang 3/29/16 3:23 PM
Deleted: trajectories

Daifeng Wang 3/29/16 3:23 PM
Deleted: meta-gene canonical temporal expression trajectories

Daifeng Wang 3/29/16 3:23 PM
Deleted: at late embryonic development

Daifeng Wang 3/29/16 3:23 PM
Deleted: that occurs early during the embryonic development,

Daifeng Wang 3/29/16 3:23 PM
Deleted: fast increasing and then unvarying pattern

Daifeng Wang 3/29/16 3:23 PM
Deleted: decaying

Daifeng Wang 3/29/16 3:23 PM
Deleted: ,

Daifeng Wang 3/29/16 3:23 PM
Deleted: an increasing pattern

Daifeng Wang 3/29/16 3:23 PM
Deleted: late

Daifeng Wang 3/29/16 3:23 PM
Deleted: two fast decaying patterns

Daifeng Wang 3/29/16 3:23 PM
Deleted: a fast

Daifeng Wang 3/29/16 3:23 PM
Deleted: at a later stage

Daifeng Wang 3/29/16 3:23 PM
Deleted: highly increasing

Daifeng Wang 3/29/16 3:23 PM
Deleted: pattern.

Daifeng Wang 3/29/16 3:23 PM
Deleted: iPDP/ePDPs

nal/external regulatory networks by the leave-one-out method; i.e., we removed one gene in the internal/external group, ran DREISS, and obtained the ordered iPDP eigenvalues for the remaining genes. We repeated the leave-one-out method for all genes, and finally found the ranges in which iPDP eigenvalues vary shown as error bars in Figure S1. We can see that the iPDP eigenvalues almost stay at the same values (small error bars) for both worm and fly, which implies that the principal dynamic patterns of worm-fly orthologous genes driven by their conserved regulatory network are robust to small changes.

Daifeng Wang 3/29/16 3:23 PM
Deleted: /ePDP

Daifeng Wang 3/29/16 3:23 PM
Deleted: /ePDP

Daifeng Wang 3/29/16 3:23 PM
Deleted: vary less than ePDP ones

Daifeng Wang 3/29/16 3:23 PM
Deleted: more

Daifeng Wang 3/29/16 3:23 PM
Deleted: than ones driven by their species-specific regulatory networks

The above results suggest that the conserved regulatory networks from orthologous meta-genes between worm and fly have similar effects to orthologous meta-genes, given their similar iPDPs (i.e., both have four patterns, as described above). The species-specific regulatory networks from species-specific meta-genes (i.e., worm-specific or fly specific TFs) have effects that differ from the orthologous meta-genes for their different ePDPs.

3.3 Orthologous genes have correlated coefficients between worm and fly for their matched internal principal dynamic patterns

In both worm and fly, we observe the similar four types of internally driven canonical temporal expression trajectories; i.e., four matched internal principal dynamic patterns (iPDPs) (Figure 3A). Thus, we are interested in seeing how individual orthologous genes relate to those dynamic patterns. We find that the worm-fly orthologous genes have correlated coefficients over each of the four iPDPs. Based on Equation (10), we can obtain the coefficients of orthologous genes for each iPDP. We find that their coefficients are significantly correlated between worm and fly iPDPs with a similar pattern (Figure 4): r=0.33 (p<2.2e-16) for the highly varied pattern at late embryonic development stages (first iPDP), r=0.66 (p<2.2e-16) for the fast decaying pattern at early embryonic development stages (second iPDP), r=0.67 (p<2.2e-16) for the slowly increasing pattern during embryonic development (third iPDP), and r=0.73 (p<2.2e-16) for the oscillation pattern during embryonic development (forth iPDP), where r represents Spearman correlation of iPDP coefficients of orthologous genes between worm and fly. This implies that, not only do the orthologous meta-genes have similar internal (conserved) regulatory effects (i.e., similar iPDPs), but the worm-fly orthologous genes also have similar internally-driven expression dynamics as resulted from their significantly correlated coefficients for iPDPs. The ePDPs between worm and fly generally do not show a high degree of matching similarity, but the worm ePDP No. 2,

Daifeng Wang 3/29/16 3:23 PM
Deleted: if we flip

Daifeng Wang 3/29/16 3:23 PM
Deleted: 3

and the fly ePDPs No. 3 are roughly representing the growing patterns. We find that orthologous gene correlation coefficients between these ePDP patterns are very small (Spearman correlation r=-0.22 of the orthologous gene coefficients of worm ePDP No.2 and fly ePDP No. 3).

3.4 Ribosomal genes have significantly larger coefficients for the internal than external principal dynamic patterns, but signaling genes exhibit the opposite trend

The ribosome produces proteins, which is an ancient process and conserved across worm and fly, organisms separated by almost a billion years of evolution. The ribosomal genes are highly expressed during embryogenesis, since intensive cell division and migration require a large amount of proteins to be synthesized. We collected 195 ribosome-related genes based on the GO annotations. We ranked the coefficients of orthologous genes for each iPDP and ePDP in ascending order, and compared the rank values of iPDP and ePDP coefficients of ribosomal genes. We found that their average ranks of iPDP coefficients are significantly larger than ePDP ones in both worm (t-test p<2.2e-16) and fly (t-test p<2.6e-13) as shown in Figure 5. This means that the ribosomal gene expression is significantly more influenced by the conserved regulatory network than by the species-specific regulatory network, which is consistent with ribosomal genes having conserved functions during embryonic development.

The orthologous genes related to signal transduction for cell-cell communication (a significantly more recent evolutionary adaptation relative to the ribosome) exhibit the opposite trend. We found that 320 signaling genes from GO annotations have significantly larger average rank values of ePDP coefficients than iPDP ones in both worm (t-test p<5.6e-11) and fly (t-test p<8.3e-4), as shown in Figure 5. This result implies that the signaling gene expression is significantly more driven by the species-specific regulatory network than by the conserved regulatory network, which is consistent with the signaling genes being commonly associated with species-specific functions, such as body plan establishment and cell differentiation.

3.5 DNA replication and Proteasome machinery are enriched in orthologous genes with high coefficients for the dynamic patterns with fast growing canonical trajectories

Daifeng Wang 3/29/16 3:23 PM
Deleted: compare with

Daifeng Wang 3/29/16 3:23 PM
Deleted: 4 and No. 5, all three of them

Daifeng Wang 3/29/16 3:23 PM
Deleted: fast decaying

Daifeng Wang 3/29/16 3:23 PM
Deleted: the

Daifeng Wang 3/29/16 3:23 PM
Deleted: =

Daifeng Wang 3/29/16 3:23 PM
Deleted: 12

Daifeng Wang 3/29/16 3:23 PM
Deleted: 3

Daifeng Wang 3/29/16 3:23 PM
Deleted: 4, and r=0.18 of worm ePDP No. 3 vs. fly ePDP No. 5

Daifeng Wang 3/29/16 3:23 PM
Deleted: , and

Daifeng Wang 3/29/16 3:23 PM
Deleted: the

Daifeng Wang 3/29/16 3:23 PM
Deleted: KS-test p<0.001) and fly (KS

Daifeng Wang 3/29/16 3:23 PM
Deleted: 5A

Daifeng Wang 3/29/16 3:23 PM
Formatted: ParaNoInd

Daifeng Wang 3/29/16 3:23 PM
Deleted: KS

Daifeng Wang 3/29/16 3:23 PM
Deleted: 7e-4

Daifeng Wang 3/29/16 3:23 PM
Deleted: KS

Daifeng Wang 3/29/16 3:23 PM
Deleted: 6e

Daifeng Wang 3/29/16 3:23 PM
Deleted: 5B

We next turn to the biological meaning of individual canonical temporal expression trajectory for iPDPs and ePDPs. For the fast-decaying pattern (2nd iPDP), we find that the DNA replication is significantly enriched in Top 300 (~10%) orthologous genes that have the most negative coefficients for this pattern, in both worm (p<1.6e-8) and fly (p<4.5e-6). The GO enrichment analysis was performed using DAVID [23]. The very negative coefficients for the fast decaying pattern mean high positive coefficients for a fast-growing pattern (vertically flipped 2nd iPDPs of worm and fly represent a fast-growing pattern), showing a drastic increase at the beginning of embryogenesis, then remain flat during the late embryogenesis (red curves in Figure 6). Most of the cell division of embryogenesis in both worm and fly happens approximately within the first 300 minutes. Then, the cell elongation and migration start to dominate the development [24,25]. The mRNA abundance of the genes involved in DNA replication may change accordingly. This is well reflected by the second iPDP. Interestingly, the original expression patterns of those top orthologous genes actually do not have fast-growing patterns (black curves in Figure 6), probably because of the combined effects of both conserved and species-specific GRN. Maternal mRNAs, which are pre-loaded before fertilization, may also mask the fast growing pattern of DNA replication genes. This pattern could only be observed after we separated the effect of two types of TFs using DREISS. In addtion, we did not find any enrichment of DNA replication in top genes of other iPDPs and ePDPs (p>0.05). Therefore, the fast-growing iPDP patterns identified by our method reveal conserved regulation on the elementary cellular process of both species (i.e. DNA replication), which should mainly be controlled by the conserved regulatory network.

Besides a fast growing pattern driven by conserved worm-fly orthologous TFs, we also identified a fast growing pattern driven by non-conserved TFs for the two species. The Top 300 orthologous genes (~10%) with the fast-growing worm ePDP (ePDP No.2) (i.e., driven by species-specific regulatory networks) are enriched in 'proteasome' (p<9.8e-16).. Protein degradation is not only a key process in apoptosis, but also throughout the entire course of development [26,27]. For example, eliminating proteins that are no longer needed is a vital process during embryo development; e.g., the maternal proteins need to be cleaned as the embryogenesis proceeds). Previous reports also showed that different species usually have different maternal mRNA in the oocyte, which indicates that species-specific strategies might be utilized to regulate the protein degradation process [28]. In this study, after separating the effect of conserved and non-conserved regulatory networks, we observed that the protein degradation is significantly enriched in the genes majorly driven by species-specific TFs in worms. In contrast, the Top 300 ortholo-

Daifeng Wang 3/29/16 3:23 PM
Deleted: and fly ePDPs

Daifeng Wang 3/29/16 3:23 PM
Deleted: shared 36 orthologous genes. 10 of them encode proteins

Daifeng Wang 3/29/16 3:23 PM
Deleted: the proteasome complex

Daifeng Wang 3/29/16 3:23 PM
Deleted: 1.2e-

Daifeng Wang 3/29/16 3:23 PM
Deleted:).

Daifeng Wang 3/29/16 3:23 PM
Deleted: . ... [1]

gous genes with fast growing fly ePDP3 are enriched in 'mitotic cell cycle' (p<3.5e-29), 'translation' (p<1e-30) and 'mitochondrion' (p<7.7e-20). Those enriched function related to energy generation is probably indicative of the large energy requirement during fly embryogenesis [30], which did not provide the evolutionary conservation of this energy-related gene regulation. Our result reveals that the fly genes associated with respiration are more up-regulated by fly-specific TFs relative to conserved TFs, and that this up-regulation evolved after the separation of worm and fly.

Besides the fast-growing pattern driven by species-specific TFs, we also observed some other interesting patterns. For example, worm ePDP3 displays a dramatic peak about 5 hours after fertilization. Among the Top 300 worm orthologous genes of this pattern, genes involved in synaptic transmission (p<5.6e-9) and cell-cell signaling (p<1e-7) are over-represented, suggesting that they are transiently activated in this stage of embryogenesis by worm-specific TFs. This observation indicates the gene regulatory network for these genes have evolved after the speciation.

3.6 Human-specific transcription factors respond to hormonal stimulation during breast cancer cell cycle

We applied DREISS to another example (also see supplement) about cancer. We are also interested to identify the gene expression dynamic patterns driven by conserved and human-specific regulatory networks during breast cancer cell cycle. Thus, we applied DREISS to a time-series gene expression data for human estrogen-responsive breast cancer cell line (ZR-75.1) before and after hormonal stimulation, which has 12 time points covering a complete mitotic cell cycle (0-32 hours) of hormonal stimulated cells [33]. The internal group, Ω is defined as a set of cross-species conserved human genes (i.e., 1132 worm-fly-human orthologs including 150 orthologous TFs), and the external group, Ψ consists of 1870 human-specific TFs. As shown in Supplemental Figure 2, the internally driven principal dynamic patterns (iPDPs) of conserved human genes include an oscillation trajectory whose period is roughly equal to a full cell cycle (iPDP No. 4), but the externally driven patterns (ePDPs No. 2-4) oscillates more frequently than internal one, which suggests that though the evolutionarily conserved TFs regulate the normal cell cycle, the human specific TFs potentially drive the abnormal cycling behaviors of conserved gene expression responding to the hormonal stimulation.

4 DISCUSSION

Daifeng Wang 3/29/16 3:23 PM
Moved down [2]: This observation indicates the gene regulatory network for these genes have evolved after the speciation.

Daifeng Wang 3/29/16 3:23 PM
Deleted: Proteins involved in calcium ion binding or GTP binding usually play a role in cell signal transduction [29]. In fact, the genes involved in Wnt signaling and MAPK signaling exhibits a two-fold change. ... [2]

Daifeng Wang 3/29/16 3:23 PM
Formatted: Font:12 pt

Daifeng Wang 3/29/16 3:23 PM
Deleted: In addition, the lack of signaling enrichment might be due to the different sampling time points. It is well-known that the Wnt signaling in worms starts as early as at the 4-cell stage, when one cell receives the signal and starts differentiation [31]. The time-series worm transcriptome data used in our study may have the resolution to detect those processes. However, since each of the first 10 cell cycles takes less than 10 minutes in the fly embryo [32], the 2 hour time interval in fly data may not have the resolution to capture the early regulatory events, such as Wnt signaling.

Daifeng Wang 3/29/16 3:23 PM
Moved (insertion) [2]

Daifeng Wang 3/29/16 3:23 PM
Formatted: Font:9 pt

Daifeng Wang 3/29/16 3:23 PM
Deleted: 1

Daifeng Wang 3/29/16 3:23 PM
Deleted: oscillation trajectory (ePDP

Daifeng Wang 3/29/16 3:23 PM
Deleted:

In this paper, we presented a novel computational method, DREISS, which decomposes time-series expression data of a group of genes into the components driven by the regulatory network inside the group (internal regulatory subsystem), and the components driven by the external regulatory network consisting of regulators outside the group (external regulatory subsystem). DREISS is a general-purpose tool that can be used to study the gene regulatory effects of any interested biological subsystems such as protein-coding transcription factors, micro-RNAs, epigenetic factors and so on. As an illustration, we applied DREISS to the time-series gene expression datasets for worm and fly embryonic developments from the modENCODE project [10], and compared the worm-fly orthologous gene expression dynamic patterns driven by the conserved regulatory network (i.e., regulation effects from orthologous TFs), with the patterns driven by the species-specific regulatory networks (i.e., regulation effects from worm or fly specific TFs). We found that the conserved TFs drive similar genomic functions, but non-conserved TFs drive species-specific functions of orthologous genes between worm and fly, implying that, in addition to having ancient conserved functions, orthologous genes have been regulated by evolutionarily younger GRNs to execute species-specific functions during the evolution. This work can be easily extended to study the regulatory effects from orthologous TFs and species-specific TFs to species-specific genes. For example, one can find the expression dynamic patterns of worm/fly specific genes driven by specific TFs, and identify the genes with strong patterns associated with worm/fly specific functions, such as body formations. To the best of our knowledge, DREISS is the first method to reveal how the evolution of GRNs affects gene expression during embryogenesis.

We emphasize that DREISS is a general-purpose method (a free downloadable R tool available from github.com/gersteinlab/dreiss). Users can define the internal group (Ω) and external group (Ψ) according to their interests. For example, if users want to identify the protein-coding expression patterns driven by miRNAs, they can define miRNAs as an external group and protein-coding genes as an internal group. Additionally, DREISS can be applied to more than two datasets, such as comparing worm, fly and human embryonic stem cell developmental data, and finding their conserved and specific developmental expression patterns. The expression patterns driven by human-specific regulatory factors will potentially help us understand human-specific developmental processes along with the associated human genes.

Due to the limited time samples in gene expression datasets, DREISS uses the simple linear state space model (i.e. the first order linear invariant difference equation) to model the temporal gene expression

dynamics, and identify principal temporal dynamic patterns. This model assumes that the gene regulatory networks controlling temporal gene expression dynamics does not change across the entire biological process such as (A, B) in Equation (1). Thus, based on the analytic analysis, the principal dynamic patterns (PDPs) must follow a small set of canonical temporal trajectories (Table 1). With the rapidly increasing gene expression data, we can extend DREISS to more advanced models such as switched and hybrid system models, non-linear models [34], which will allow us to study the gene regulatory networks are time varying, and potentially find the more temporal gene expression patterns capturing the more complex gene regulatory activities.

FIGURE CAPTIONS

Figure 1 DREISS workflow. **1:** DREISS models temporal gene expression dynamics using state-space models in control theory. The "state" refers to the expressions for a large group of genes of interest, such as the worm-fly orthologous genes investigated here. The "control" refers to any other group of genes that contribute to gene expressions of the "state", such as the species-specific TF studied here. **2**: it then projects high-dimensional gene expression space to lower-dimensional meta-gene expression spaces using dimensionality reduction techniques. **3**: it derives the effective state-space models for meta-genes so that model parameters can be estimated. **4**: it then identifies the meta-gene expression dynamic patterns; i.e., canonical temporal expression trajectories driven by "state" (internal) and by "control" (external) based on the analytic solutions to estimated models. **5**: it finally calculates the coefficients of genes for the dynamic patterns of linear transformations between genes and meta-genes.

Figure 2 State space model for genes and the effective model for meta-genes. A) linear state space model for a given subsystem's gene expression; i.e., linear first-order difference equations in Equation (1), is used to formulate temporal gene expression dynamics for a given subsystem, the internal group Ω (comprising N_1 genes) with external regulations from the external group Ψ (comprising N_2 genes) at time points 1, 2, … , T. The vector $X_t \in \Re^{N_1 \times 1}$, the "state", includes N_1 gene expression levels at time t in Ω, and the vector $U_t \in \Re^{N_2 \times 1}$, the "input or control", includes N_2 gene expression levels at time t in Ψ. The system matrix $A \in \Re^{N_1 \times N_1}$ captures internal causal interactions among genes in Ω (i.e., the i^{th}, j^{th} element of A, A_{ij} describes the contribution from the j^{th} gene expression at time t to the i^{th} gene expression at the next time t+1). The control matrix $B \in \Re^{N_1 \times N_2}$ captures external causal regulations from the genes in Ψ to genes in Ω (i.e., the i^{th}, j^{th} element of B, B_{ij} describes the contribution from the j^{th} gene expression in Ψ at time t to the i^{th} gene expression in Ω at the next time t+1). **B)** Meta-gene expression levels. The meta-gene expression levels are obtained by $\tilde{X}_t = W_X^* X_t; \tilde{U}_t = W_U^* U_t$, where $\tilde{X}_t \in \Re^{M_1 \times 1}$, the "meta-gene state", includes M_1 ($<< N_1$ and $<T$) meta-gene expression levels; i.e., the first M_1 elements of the t^{th}

Daifeng Wang 3/29/16 3:23 PM
Deleted: 2

row of the matrix whose columns are right-singular vectors of the matrix $[X_1 X_2 ... X_T]$ in Ω by the singular value decomposition (SVD) [19]; the vector $\tilde{U}_t \in \Re^{M_2 \times 1}$, the "meta-gene input or control", includes M_2 (<< N_2 and <T) meta-gene expression levels (i.e., the first M_2 elements of the t^{th} row of the matrix whose columns are right-singular vectors of the matrix SVD of matrix $[U_1 U_2 ... U_T]$ at time t in Ψ; $W_X \in \Re^{N_1 \times M_1}$ is the linear projection matrix of SVD from M_1 meta-gene expression space to N_1 gene expression space in X, $W_U \in \Re^{N_2 \times M_2}$ is the linear projection matrix of SVD from M_2 meta-gene expression space to N_2 gene expression space in Ψ), and $(.)^*$ is a pseudo-inverse operation; i.e., $W^* W = I$, where I is the identity matrix. **C)** Effective state space model for meta-genes. The effective state-space model for meta-genes, Equation (4) is obtained by using linear projections W_X and W_U between genes and meta-genes from Equations (1-3). The effective meta-gene system matrix $\tilde{A} = W_X^* A W_X \in \Re^{M_1 \times M_1}$ captures internal causal interactions among meta-genes in Ω (i.e., the i^{th}, j^{th} element of $\tilde{A}$ ($\tilde{A}_{ij}$) describes the contribution from the j^{th} meta-gene expression at time t to i^{th} meta-gene expression at next time t+1), and the effective control matrix $\tilde{B} = W_X^* B W_U \in \Re^{M_1 \times M_2}$ captures external causal regulations from meta-genes in Ψ to meta-genes in Ω (i.e., the i^{th}, j^{th} element of $\tilde{B}$, $\tilde{B}_{ij}$ describes the contribution from the j^{th} meta-gene expression in Ψ at time t to i^{th} meta-gene expression in Ω at next time t+1). Equation (4) describes the effective state space model for the meta-genes in Ω, whose expression dynamics are determined by $\tilde{A}$ and $\tilde{B}$. Because the meta-gene dimension, M_1 (M_2) is less than T, and much less than N_1 (N_2), we can estimate $\tilde{A}$ and $\tilde{B}$

Figure 3 Principal dynamic patterns of orthologous genes between worm and fly during embryonic development. A) Metagenes of orthologous genes have similar internal driven principal dynamic patterns. Meta-gene canonical temporal expression trajectories driven by conserved regulatory networks (i.e., internal principal dynamic patterns, iPDPs) include four major patterns in both worm and fly embryonic development: 1) a highly varied pattern late (iPDP with the real eigenvalue No. 1); 2) a fast decaying pattern early (iPDP with the real eigenvalue No. 2); 3) a slowly increasing pattern (iPDP with the real eigenvalue No. 3); and 4) an oscillating pattern (iPDP with the complex eigenvalue). **B)** Metagenes of orthologous genes have different external driven principal dynamic patterns. Worm and fly have very different external principal dynamic patterns (ePDPs); i.e., the patterns driven by species-specific TFs. The meta-gene dynamic patterns driven by the worm-specific regulatory network; i.e., worm ePDPs consist of a varied pattern that decreases until the middle stage and then increases (ePDP No.1), an increasing pattern (ePDP No.2), a varied pattern with a peak entering middle stage ((ePDP No.3), a pattern that varies early and then increases during the embryonic development (ePDP No.4), and a cosine-like oscillating pattern with roughly two periods during the embryonic development (ePDP No.5). The fly ePDPs, however, have a

Daifeng Wang 3/29/16 3:23 PM
Deleted: as follows

Daifeng Wang 3/29/16 3:23 PM
Deleted: canonical temporal expression trajectories

Daifeng Wang 3/29/16 3:23 PM
Deleted: at late

Daifeng Wang 3/29/16 3:23 PM
Deleted: real eigenvalue

Daifeng Wang 3/29/16 3:23 PM
Deleted: 1),

Daifeng Wang 3/29/16 3:23 PM
Deleted: varied

Daifeng Wang 3/29/16 3:23 PM
Deleted: at early

Daifeng Wang 3/29/16 3:23 PM
Deleted: real eigenvalue No. 2), a fast increasing and then unvarying pattern (real eigenvalue No. 3), a decaying pattern (real eigenvalue No. 4), and an increasing pattern at late embryonic development (real eigenvalue

Daifeng Wang 3/29/16 3:23 PM
Deleted:

Daifeng Wang 3/29/16 3:23 PM
Deleted: two fast decaying patterns at early

varied pattern with low expression at the early stage (ePDP No.1), a sine-like oscillating pattern with roughly one period during the embryonic development (ePDP No.2), an increasing pattern (ePDP No.3), another sine-lie oscillating pattern with roughly two periods during the embryonic development (ePDP No.4), and a varied pattern that is like damped oscillation (ePDP No.5).

Figure 4 Orthologous genes have correlated coefficients between worm and fly for their matched internal principal dynamic patterns. The worm-fly orthologous genes have correlated coefficients over each of four iPDPs. Their coefficients are significantly correlated between worm and fly iPDPs with a similar pattern: r=0.33 (p<2.2e-16) for the highly varied pattern at late embryonic development (first iPDP), r=0.66 (p<2.2e-16) for the fast decaying pattern at early embryonic development (second iPDP), r=0.67 (p<2.2e-16) for the slowly increasing pattern during embryonic development (third iPDP), and r=0.73 (p<2.2e-16) for the oscillation pattern during embryonic development (forth iPDP).

Figure 5 Ribosomal genes have significantly larger coefficients for internal than external principal dynamic patterns, but signaling genes exhibit the opposite trend. The rank values in ascending order of iPDP and ePDP coefficients of ribosomal and signaling genes (cell-cell communication) genes are compared. The y-axis shows the distributions of rank values. Ribosomal genes (white boxes): their average rank values of iPDP coefficients are significantly larger than ePDP ones in both worm (t-test p<2.2e-16) and fly (t-test p<5.6e-11). Signaling genes (grey boxes): they have significantly larger average rank values of ePDP coefficients than iPDP ones in both worm (t-test p<2.6e-13) and fly (t-test p<8.3e-4).

Figure 6 DNA replication is enriched in orthologous genes with high coefficients for the dynamic patterns with fast growing canonical trajectories. For the fast-decaying pattern (2nd iPDP), we found that the DNA replication is significantly enriched in Top 300 (~10%) orthologous genes that have the most negative coefficients for this pattern, in both worm (p<1.6e-8) and fly (p<4.5e-6). The very negative coefficients for the fast decaying pattern means high positive coefficients for a fast-growing pattern, showing a drastic increase at the beginning of embryogenesis, then remain flat during the late embryogenesis (red curves). The original expression patterns of those top orthologous genes actually do not have fast-growing patterns (black curves).

Figure S1 Principal dynamic patterns and their eigenvalues. Internal principal dynamic patterns (iPDPs) of orthologs during worm and fly embryonic development. Barplots show the eigenvalues of iPDPs. The error bar for each eigenvalue tells the its variation range. We left one gene out, and calculated eigenvalues for the remaining genes thus obtaining the eigenvalue variations. The curves show the canonical temporal expression trajectories of iPDPs.

Daifeng Wang 3/29/16 3:23 PM
Deleted: real eigenvalue No. 1 and 2), a fast increasing pattern at late embryonic dev ... [3]

Daifeng Wang 3/29/16 3:23 PM
Deleted: highly increasing

Daifeng Wang 3/29/16 3:23 PM
Deleted: pattern (complex eigenvalue

Daifeng Wang 3/29/16 3:23 PM
Formatted ... [4]

Daifeng Wang 3/29/16 3:23 PM
Formatted ... [5]

Daifeng Wang 3/29/16 3:23 PM
Formatted ... [6]

Daifeng Wang 3/29/16 3:23 PM
Formatted ... [7]

Daifeng Wang 3/29/16 3:23 PM
Deleted: The

Daifeng Wang 3/29/16 3:23 PM
Deleted: : the

Daifeng Wang 3/29/16 3:23 PM
Deleted: KS-test p<0.001) and fly (KS

Daifeng Wang 3/29/16 3:23 PM
Formatted ... [8]

Daifeng Wang 3/29/16 3:23 PM
Deleted:). The iPDP

Daifeng Wang 3/29/16 3:23 PM
Deleted: ePDP coefficients of signaling

Daifeng Wang 3/29/16 3:23 PM
Deleted: cell-cell communication) are ... [9]

Daifeng Wang 3/29/16 3:23 PM
Deleted: KS

Daifeng Wang 3/29/16 3:23 PM
Deleted: 7e-4

Daifeng Wang 3/29/16 3:23 PM
Deleted: KS

Daifeng Wang 3/29/16 3:23 PM
Deleted: 6e-4). The boxplots show t ... [10]

Daifeng Wang 3/29/16 3:23 PM
Formatted ... [11]

Daifeng Wang 3/29/16 3:23 PM
Formatted ... [12]

Daifeng Wang 3/29/16 3:23 PM
Deleted: **A)** internal

Daifeng Wang 3/29/16 3:23 PM
Deleted: PDPs); **B)** external PDPs

Daifeng Wang 3/29/16 3:23 PM
Deleted: PDPs

Daifeng Wang 3/29/16 3:23 PM
Deleted: PDPs

Figure S2 Internally and externally principal dynamic patterns of cross-species conserved gene expression during human breast cancer cell cycle after hormonal stimulation. The horizontal axis represents 12 time points from 0 to 32 hours during a complete mitotic breast cancer cell cycle (E-TABM-631, ArrayExpress). The vertical axis represents the normalized PDP expression with the vector norm equal to one. The internal group is defined as a set of cross-species conserved human genes (i.e., 1132 worm-fly-human orthologs; including 150 orthologous TFs), and the external group consists of 1870 human-specific TFs.

Table S1 Examples of internal and external regulatory networks.

REFERENCES

1. Kim PM, Tidor B (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res 13: 1706-1718.
2. Vilar JM (2006) Modularizing gene regulation. Mol Syst Biol 2: 2006 0016.
3. Peter IS, Davidson EH (2011) Evolution of gene regulatory networks controlling body plan development. Cell 144: 970-985.
4. Chen K, Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet 8: 93-103.
5. Levin M, Hashimshony T, Wagner F, Yanai I (2012) Developmental milestones punctuate gene expression in the Caenorhabditis embryo. Dev Cell 22: 1101-1108.
6. Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, et al. (2010) Gene expression divergence recapitulates the developmental hourglass model. Nature 468: 811-814.
7. Yanai I, Peshkin L, Jorgensen P, Kirschner MW (2011) Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility. Dev Cell 20: 483-496.
8. Irie N, Kuratani S (2011) Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. Nat Commun 2: 248.
9. Casci T (2011) Development: Hourglass theory gets molecular approval. Nat Rev Genet 12: 76.
10. Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, et al. (2014) Comparative analysis of the transcriptome across distant species. Nature 512: 445-448.
11. Brogan WL (1991) Modern control theory. Englewood Cliffs, N.J.: Prentice Hall. xviii, 653 p. p.
12. Rangel C, Angus J, Ghahramani Z, Lioumi M, Sotheran E, et al. (2004) Modeling T-cell activation using gene expression profiling and state-space models. Bioinformatics 20: 1361-1372.
13. Bansal M, Della Gatta G, di Bernardo D (2006) Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics 22: 815-822.
14. Huang S, Ingber DE (2006) A non-genetic basis for cancer progression and metastasis: self-organizing attractors in cell regulatory networks. Breast Dis 26: 27-54.
15. Saeys Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23: 2507-2517.
16. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, et al. (1998) The transcriptional program of sporulation in budding yeast. Science 282: 699-705.
17. Wang D, Arapostathis A, Wilke CO, Markey MK (2012) Principal-oscillation-pattern analysis of gene expression. PLoS One 7: e28805.
18. Wang D, Markey MK, Wilke CO, Arapostathis A (2012) Eigen-genomic system dynamic-pattern analysis (ESDA): modeling mRNA degradation and self-regulation. IEEE/ACM Trans Comput Biol Bioinform 9: 430-437.
19. Golub GH, Van Loan CF (1996) Matrix computations. Baltimore: Johns Hopkins University Press. xxvii, 694 p. p.
20. Cull P, Flahive ME, Robson RO (2005) Difference equations : from rabbits to chaos. New York: Springer. xiii, 392 p. p.
21. Reece-Hoyes JS, Deplancke B, Shingles J, Grove CA, Hope IA, et al. (2005) A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks. Genome Biol 6: R110.
22. Shazman S, Lee H, Socol Y, Mann RS, Honig B (2014) OnTheFly: a database of Drosophila melanogaster transcription factors and their binding sites. Nucleic Acids Res 42: D167-171.

23. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.
24. Bate M, Martinez Arias A (1993) The Development of Drosophila melanogaster. Plainview, N.Y.: Cold Spring Harbor Laboratory Press.
25. Baugh LR, Hill AA, Slonim DK, Brown EL, Hunter CP (2003) Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. Development 130: 889-900.
26. DeRenzo C, Seydoux G (2004) A clean start: degradation of maternal proteins at the oocyte-to-embryo transition. Trends Cell Biol 14: 420-426.
27. Du Z, He F, Yu Z, Bowerman B, Bao Z (2015) E3 ubiquitin ligases promote progression of differentiation during C. elegans embryogenesis. Dev Biol 398: 267-279.
28. Shen-Orr SS, Pilpel Y, Hunter CP (2010) Composition and regulation of maternal and zygotic transcriptomes reflects species-specific reproductive mode. Genome Biol 11: R58.
29. Aspenstrom P (2004) Integration of signalling pathways regulated by small GTPases and calcium. Biochim Biophys Acta 1742: 51-58.
30. Tennessen JM, Bertagnolli NM, Evans J, Sieber MH, Cox J, et al. (2014) Coordinated metabolic transitions during Drosophila embryogenesis and the onset of aerobic glycolysis. G3 (Bethesda) 4: 839-850.
31. Sawa H, Korswagen HC (2013) Wnt signaling in C. elegans. WormBook: 1-30.
32. Gilbert SF (2000) Developmental biology. Sunderland, Mass.: Sinauer Associates. xviii, 749 p. p.
33. Mutarelli M, Cicatiello L, Ferraro L, Grober OM, Ravo M, et al. (2008) Time-course analysis of genome-wide gene expression data from hormone-responsive human breast cancer cells. BMC Bioinformatics 9 Suppl 2: S12.
34. Schaft AJvd, Schumacher JM (2000) An introduction to hybrid dynamical systems. London ; New York: Springer. xi, 174 p. p.

Page 15: [1] Deleted Daifeng Wang 3/29/16 3:23 PM

.

Besides the 36 shared genes in the fast-growing pattern driven by species-specific TFs, we also observed some other interesting results. Among the Top 300 worm orthologous genes with fast-growing ePDPs, genes involved in calcium ion binding (p<2e-6), GTP binding (p<7e-3) and neuron differentiation (p<0.05) are over-represented, suggesting that they are activated in the early stage of embryogenesis by worm-specific TFs.

Page 16: [2] Deleted Daifeng Wang 3/29/16 3:23 PM

Proteins involved in calcium ion binding or GTP binding usually play a role in cell signal transduction [29]. In fact, the genes involved in Wnt signaling and MAPK signaling exhibits a two-fold change.

In contrast, the Top 300 fly genes with a fast-growing ePDP show no enrichment in signaling transduction or cell differentiation. Instead, functions associated with respiration, such as oxidative phosphorylation, are enriched (p<5e-10). The enrichment of energy generation in the Top 300 fly genes with a fast-growing ePDP

Page 20: [3] Deleted Daifeng Wang 3/29/16 3:23 PM

real eigenvalue No. 1 and 2), a fast increasing pattern at late embryonic development (real eigenvalue No. 3),

Page 20: [4] Formatted Daifeng Wang 3/29/16 3:23 PM

Font:Italic

Page 20: [5] Formatted Daifeng Wang 3/29/16 3:23 PM

Font:Italic

Page 20: [6] Formatted Daifeng Wang 3/29/16 3:23 PM

Font:Italic

Page 20: [7] Formatted Daifeng Wang 3/29/16 3:23 PM

Font:Italic

Page 20: [8] Formatted Daifeng Wang 3/29/16 3:23 PM

Font:Italic

Page 20: [9] Deleted Daifeng Wang 3/29/16 3:23 PM

cell-cell communication) are compared:

Page 20: [10] Deleted	Daifeng Wang	3/29/16 3:23 PM

6e-4). The boxplots show the iPDP and ePDP coefficients of ribosomal related genes in fly.

Page 20: [11] Formatted	Daifeng Wang	3/29/16 3:23 PM

Font:Italic

Page 20: [12] Formatted	Daifeng Wang	3/29/16 3:23 PM

Font:Italic