DREISS: Decomposition of gene Regulatory network into External and Internal components based on State Space models

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]Brookhaven National Laboratory, Upton, NY, USA. *Correspondence to: pi@gersteinlab.org

ABSTRACT

Motivation: Gene regulatory network at the system level can be modularized into multiple interconnected subsystems. A subsystem's gene expression may be controlled by its internal regulatory factors only, or also by other external subsystems. How to distinguish the internal and external regulatory effects on subsystems and decide their self-regulatory degrees become interesting.

Results: Here, we developed a novel computational method termed DRIESS (Decomposition of gene Regulatory network into External and Internal components based on State Space models), which integrates a state space model and dimensionality reduction analysis for a given subsystem to dissect the effects of other regulatory subsystems on its gene expression. Here, we focused on the evolutionary conserved subsystem consisting of orthologous genes, and applied DREISS to time-series gene expression datasets for worm and fly embryonic development from the modENCODE project. DREISS derived and estimated the effective state space model for meta-genes comprising a low-dimensional space explaining the most gene expression covariance, and then found the gene expression dynamics via analytic analysis to the effective model. We found that between the two species, the expression patterns of orthologous genes driven by conserved regulatory subsystems (i.e., orthologous TFs) are more similar to each other than those driven by species-specific regulatory subsystems. Moreover, genes involved in evolutionarily ancient functions (e.g. ribosome), and more recently evolved functions (e.g., cell-cell communication) are significantly more controlled by the conserved and species-specific regulatory subsystems, respectively. Our results imply that the expression of ancient genes for the conserved developmental functions is tightly controlled, as compared to the ones for the species-specific functions.

Conclusion: We demonstrated the capabilities of DREISS for studying the effects of different evolutionary gene regulatory subsystems on gene expression during the embryogenesis across distant species. Despite striking morphological differences, some basic developmental processes are still tightly under the control of ancient transcriptional regulatory factors. DREISS can be also used as a general-

*To whom correspondence should be addressed.

Daifeng Wang 1/11/15 5:46 PM
Deleted: Motivation: It is well known that the evolution of gene expression is more important than the evolution of gene sequences in determining developmental morphologies. Comparative transcriptomics provides a general approach to study the evolution of gene expression (i.e., by elucidating the regulatory mechanisms from different evolutionary gene regulatory networks that control the gene expression in evolution). However, current strategies, such as directly comparing developmental expression profiles, do not allow us to understand the mechanisms due to differences in sample size and a lack of synchronization.

Daifeng Wang 1/11/15 5:46 PM
Deleted:) to address this problem. DRIESS

Daifeng Wang 1/11/15 5:46 PM
Deleted: the

Daifeng Wang 1/11/15 5:46 PM
Deleted: in engineering

Daifeng Wang 1/11/15 5:46 PM
Deleted: evolutionary

Daifeng Wang 1/11/15 5:46 PM
Deleted: networks

Daifeng Wang 1/11/15 5:46 PM
Deleted: We

Daifeng Wang 1/11/15 5:46 PM
Deleted: developmental data in

Daifeng Wang 1/11/15 5:46 PM
Deleted: networks

Daifeng Wang 1/11/15 5:46 PM
Deleted: ones.

Daifeng Wang 1/11/15 5:46 PM
Deleted: -

Daifeng Wang 1/11/15 5:46 PM
Deleted: networks

Daifeng Wang 1/11/15 5:46 PM
Deleted: evolution of

purpose tool (github.com/gersteinlab/dreiss) to study and compare the effects from different types of regulatory subsystems, such as TFs vs. miRNAs across various datasets.

1 INTRODUCTION

Developmental gene regulatory networks (GRNs) control gene expression to determine developmental processes. These GRNs have been evolved, and thus are difficult to understand their regulatory mechanisms at the system level. Thus, one typically compares developmental gene expression across species to infer activities of developmental GRNs. For example, embryogenesis provides a platform to study the evolution of gene expression between different species. Recent work showed that significant biological insight can be gained by cross-species comparisons of the expression profiles during embryogenesis for worms (Levin, et al., 2012), flies (Kalinka, et al., 2010), frogs (Yanai, et al., 2011) and several other vertebrates (Irie and Kuratani, 2011). For example, it was found that the hourglass patterns of orthologous genes have minimal expression divergence across species within the same phylum during the phylotypic stage at embryonic development. In addition, the conserved hourglass patterns were observed within a single species while comparing the developmental gene expression data across distant species, such as worm and fly (Gerstein, et al., 2014); i.e., the expression divergence among evolutionarily conserved genes become minimal during the phylotypic stage in both worm and fly. However, how those orthologous genes in each species eventually contribute to their species-specific phenotypes is less studied due to the lack of appropriate computational approaches.

Gene regulatory networks are highly modular, and consist of various sub-networks, each of which includes a number of regulatory factors representing a regulatory subsystem to drive particular gene regulatory functions (Kim and Tidor, 2003; Vilar, 2006). Those subsystems interact with each other, 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 comprising complex evolutionary GRNs (Peter and Davidson, 2011). Those regulatory subsystems have driven very diverse developmental functions, from the highly conserved (e.g. DNA replication) to the species-specific (e.g. body segmentation). Unfortunately, existing experimental gene expression data cannot decouple the expression components that are driven by different subsystems. For example, the orthologous genes can be regulated by orthologous and species-specific transcription factors, which comprise a regulatory network that is "internal" and "external" regulatory subsystem to orthologous genes, respectively.

Daifeng Wang 1/11/15 5:46 PM
Deleted: networks

Daifeng Wang 1/11/15 5:46 PM
Deleted: Evolution has rendered these

Daifeng Wang 1/11/15 5:46 PM
Deleted: too complex

Daifeng Wang 1/11/15 5:46 PM
Deleted: fully

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Levin, et al., 2012)

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Kalinka, et al., 2010)

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Yanai, et al., 2011)

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Irie and Kuratani, 2011)

Daifeng Wang 1/11/15 5:46 PM
Deleted: similar

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Gerstein, et al., 2014)

Daifeng Wang 1/11/15 5:46 PM
Deleted: The

Daifeng Wang 1/11/15 5:46 PM
Deleted: factors

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Peter and Davidson, 2011)

Daifeng Wang 1/11/15 5:46 PM
Deleted: factors

Daifeng Wang 1/11/15 5:46 PM
Deleted: factors

Thus, we need computational methods to derive the contribution from each factor or subsystem from the gene expression data. In this study, we propose a novel computational method, DREISS - Decomposition of gene Regulatory network into External and Internal components based on State Space models. We identify temporal gene expression dynamic patterns for evolutionary conserved genes during embryonic development, as driven by conserved and species-specific regulatory subsystems, thereby advancing our current understanding of GRNs in evolution and differentiation during development.

The state-space model has long been widely used in engineering (Brogan, 1991) and analyzing gene expression dynamics (Bansal, et al., 2006; Huang and Ingber, 2006; Rangel, et al., 2004). It models the dynamical system output as a function of both current internal system state and external input signal. Similarly, for the orthologous genes, their expression at the next developmental stage can be predicted from their expression (internal) and species-specific regulatory factors (external) at the current stage. Unlike previous work that calculates the expression correlation between ~~individual~~ genes, the state-space model predicts the causal relationships at the system level. Previous work applied the state-space model to study the gene expression dynamics focused 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 also 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 (Chu, et al., 1998; Kim and Tidor, 2003; Saeys, et al., 2007)) as well as the principal dynamic patterns for those meta-genes (Wang, et al., 2012; Wang, et al., 2012). DREISS applies the dimensionality reduction to the gene expression data, develops an effective state-space model for their meta-genes, and identifies the dynamic patterns driven by effective conserved and species-specific meta-gene regulatory networks according to the model's analytic characteristics. Those dynamic patterns reveal temporal gene expression components that are controlled by conserved or species-specific GRNs.

We applied DREISS to the gene expression data during embryonic development for two model organisms, worm (*Caenorhabditis elegans*) and fly (*Drosophila melanogaster*), and in both species, we identified the expression patterns of worm-fly orthologs driven by the conserved regulatory network consist-

Daifeng Wang 1/11/15 5:46 PM
Deleted: GRN

Daifeng Wang 1/11/15 5:46 PM
Deleted: networks

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Brogan, 1991)

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Bansal, et al., 2006; Huang and Ingber, 2006; Rangel, et al., 2004)

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Kim and Tidor, 2003; Saeys, et al., 2007)

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Wang, et al., 2012; Wang, et al., 2012)

ing of the worm-fly orthologous transcription factors (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 been regulated by species-specific TFs to involve in species-specific developmental processes. In summary, DRIESS provides a framework to analyze distantly and closely related species to understand the gene regulatory mechanisms during development. Moreover, by relating this to human genes, such analyses may inform our understanding of development as it relates to humans, such as developmental in the human brain and pathologies such as Alzheimer's disease.

Daifeng Wang 1/11/15 5:46 PM
Deleted: .

2 MATERIALS AND METHODS

DREISS consists of five major steps (Figure 1):

Daifeng Wang 1/11/15 5:46 PM
Deleted:).

Step 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.

Step 2: Due to the limited number of temporal samples in gene expression experiments, we do not have enough data to estimate the parameters of the state-space models that capture interactions between hundreds of genes. Therefore, DREISS projects high-dimensional gene expression space to lower-dimensional meta-gene expression spaces using dimensionality reduction techniques.

Step 3: DREISS then derives the effective state-space models for meta-genes so that model parameters can be estimated.

Step 4: DREISS then identifies the meta-gene expression dynamic patterns driven by "state" (internal) and by "control" (external) based on the analytic solutions to estimated models.

Daifeng Wang 1/11/15 5:46 PM
Deleted: Finally, in

Step 5, we calculate the coefficients of genes for those dynamic patterns for linear transformations between genes and meta-genes. Moreover, DREISS allows us to compare the dynamic expression patterns of multiple datasets with samples taken at different times. We describe DREISS in detail in each step as follows.

2.1 State-space models for temporal gene expression dynamics

A gene regulatory system is made up of various subsystems (Kim and Tidor, 2003; Vilar, 2006). Those subsystems work together rather than individually to execute the regulatory functions. Given a group of N_1 genes in a subsystem, their gene expression levels (X) are not only controlled by internal interactions among X, but also affected by the regulatory factors from other subsystems outside X (external regulations). For example, in our study, we consider the worm-fly orthologous genes as the X group. The worm-fly orthologous TFs from the X group are the internal regulatory factors, and non-orthologous TFs such as worm- or fly- specific TFs are the external regulatory factors to the X group. Both 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 future times). Thus, the regulatory mechanisms for the gene expressions form a control system. In this study, we used a state-space model (linear first-order difference equations, Figure 2), which has been commonly used in control engineering, to formulate temporal gene expression dynamics for the gene group X (comprising N_1 genes) with external regulations from the gene group U (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 group X, 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 group U. The system matrix $A \in \Re^{N_1 \times N_1}$ captures internal causal interactions among genes in X (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 U to genes in X (i.e., the i^{th}, j^{th} element of B, B_{ij} describes the contribution from the j^{th} gene expression in U at time t to the i^{th} gene expression in X at the next time t+1). According to the state-space model (1), the gene expression dynamics in X are determined by system matrix A and the control matrix B.

2.2 Dimensionality reduction from genes to meta-genes

However, 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 samples needed to estimate the large matrices A and B when X and U have hundreds or thousands of genes. Thus, we project high dimensional temporal gene expressions to much lower dimensional meta-gene expression levels using dimensionality reduction (Figure 3). Those meta-gene expression levels should capture original gene expression

Daifeng Wang 1/11/15 5:46 PM
Deleted: Given a group of N_1 genes

Daifeng Wang 1/11/15 5:46 PM
Deleted: other

Daifeng Wang 1/11/15 5:46 PM
Deleted: $X_{t+1} = AX_t + BU_t$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $A \in \Re^{N_1 \times N_1}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $B \in \Re^{N_1 \times N_2}$

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", includes M_1 (<< N_1 and <T) meta-gene expression levels; i.e., the values of first M_1 singular vectors from singular value decomposition (SVD) of matrix $[X_1 X_2 ... X_T]$ at time t in group X; the vector $\tilde{U}_t \in \Re^{M_2 \times 1}$, the "meta-gene input or control", include M_2 (<< N_2 and <T) meta-gene expression levels (i.e., the values of the first M_2 singular vectors of SVD of matrix $[U_1 U_2 ... U_T]$ at time t in group U; $W_X \in \Re^{N_1 \times M_1}$ 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 U), and $(.)^*$ is a pseudo-inverse operation; i.e., $W^* W = I$, where I is the identity matrix.

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

Next, we can 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 4). 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)

, and by 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),

$\tilde{X}_{t+1} = \underbrace{W_X^* A W_X}_{\tilde{A}} \tilde{X}_t + \underbrace{W_X^* B W_U}_{\tilde{B}} \tilde{U}_t \Rightarrow \tilde{X}_{t+1} = \tilde{A}\tilde{X}_t + \tilde{B}\tilde{U}_t$ (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 X (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 U to meta-genes in X (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 U at time t to i^{th} meta-gene expression in X at next time t+1). Equation (4) describes the effective state space model for the meta-genes in X, whose expression dynamics are determined by $\tilde{A}$ and $\tilde{B}$. Because the me-

Daifeng Wang 1/11/15 5:46 PM
Deleted: $X_t = W_X \tilde{X}_t; U_t = W_U \tilde{U}_t$

Daifeng Wang 1/11/15 5:46 PM
Formatted: Font:12 pt

Daifeng Wang 1/11/15 5:46 PM
Deleted: $[X_1 X_2 ... X_T]$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $[U_1 U_2 ... U_T]$

Daifeng Wang 1/11/15 5:46 PM
Deleted: and

Daifeng Wang 1/11/15 5:46 PM
Deleted:).

Daifeng Wang 1/11/15 5:46 PM
Deleted: $W_X \tilde{X}_{t+1} = A W_X \tilde{X}_t + B W_U \tilde{U}_t$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $W_X^* \in \Re^{M_1 \times N_1}$ s.t. $W_X^* W_X = I$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{X}_{t+1} = \underbrace{W_X^* A W_X}_{\tilde{A}} \tilde{X}_t + \underbrace{W_X^* B W_U}_{\tilde{B}}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{A} = W_X^* A W_X \in \Re^{M_1 \times M_1}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{A}$ ($\tilde{A}_{ij}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{B} = W_X^* B W_U \in \Re^{M_1 \times M_2}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{B}$, $\tilde{B}_{ij}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{A}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{B}$

ta-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:

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

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

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

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

The effective internal system matrix $\tilde{A}$ and external control matrix $\tilde{B}$ can be estimated by:

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

, where $\Upsilon^* \in \Re^{(T-1)\times(M_1+M_2)}$ is the pseudo-inverse of Υ; i.e.

$\Upsilon\Upsilon^* = I$, with $M_1 < N_1, M_2 < N_2, M_1 + M_2 < T, t = 1,2,...,T$.

2.4 Identification of internally and externally driven meta-gene expression dynamic patterns

According to the analytic solution to Equation (4), the components of meta-gene expressions in X driven by effective internal regulations ($\tilde{X}^I_{t+1} = \tilde{A}\tilde{X}^I_t$) are linear combinations of M_1 dynamic patterns determined by the eigenvalues of the effective system matrix $\tilde{A}$ as follows:

$$\tilde{X}^I_t = \sum_{p=1}^{M_1} \lambda_p^t \tilde{V}_p^A \quad (8)$$

, where $\lambda_p(\tilde{V}_p^A)$ is the p^{th} eigenvalue (eigenvector) of $\tilde{A}$, which determines the p^{th} dynamic pattern driven by effective internal regulations, defined as the p^{th} internal principal dynamic pattern (iPDP) = $\left[\lambda_p^1\ \lambda_p^2\ \ldots\ \lambda_p^T\right]$. If an eigenvalue λ is complex when is $\tilde{A}$ asymmetric, then its conjugate $\bar{\lambda}$ is also an eigen-

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{A}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{B}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{X}_{t+1} = \tilde{A}\tilde{X}_t + \tilde{B}\tilde{U}_t = \begin{bmatrix} \tilde{A} & \tilde{B}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\underbrace{\begin{bmatrix} \tilde{X}_2 & \tilde{X}_3 & \cdots & \tilde{X}_T \end{bmatrix}}_{\mathrm{Z}} = \begin{bmatrix}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\mathrm{Z} \in \Re^{M_1\times(T-1)}$ and $\Upsilon \in \Re^{(M_1}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{A}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{B}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\begin{bmatrix} \tilde{A} & \tilde{B} \end{bmatrix} = \mathrm{Z}\Upsilon^*$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\Upsilon^* \in \Re^{(T-1)\times(M_1+M_2)}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: Υ

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\Upsilon\Upsilon^* = I$, with $M_1 < N_1, M_2$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{A}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{X}^I_t = \sum_{p=1}^{M_1} \lambda_p^t \tilde{V}_p^A$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\lambda_p(\tilde{V}_p^A)$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{A}$

value, so we sum its iPDP and its conjugate eigenvalue, $\bar{\lambda}$'s iPDP as a unified iPDP with real elements equal to $\left[\lambda_p^1+\bar{\lambda}_p^1\ \lambda_p^2+\bar{\lambda}_p^2\ \ldots\ \lambda_p^T+\bar{\lambda}_p^T\right]$. Similarly, the components of meta-gene expressions in X driven by effective external regulations from U, i.e., $\tilde{X}_{t+1}^E = \tilde{B}\tilde{X}_t^E$ are linear combinations of M_2 dynamic patterns determined by the eigenvalues of the effective system matrix $\tilde{B}$ as follows:

$$\tilde{X}_t^E = \sum_{q=1}^{M_2} \sigma_q^t \tilde{V}_q^B \qquad (9)$$

, where $\sigma_q(\tilde{V}_q^B)$ is the q^{th} eigenvalue(eigenvector) of $\tilde{B}$, which determines q^{th} dynamic pattern driven by effective external regulations, defined as q^{th} external principal dynamic pattern (ePDP) = $\left[\sigma_q^1\ \sigma_q^2\ \ldots\ \sigma_q^T\right]$. If an eigenvalue σ is complex, then its conjugate $\bar{\sigma}$ is also an eigenvalue, so we sum its ePDP and its conjugate eigenvalue, $\bar{\sigma}$'s ePDP as a unified ePDP with real elements equal to $\left[\sigma_p^1+\bar{\sigma}_p^1\ \sigma_p^2+\bar{\sigma}_p^2\ \ldots\ \sigma_p^T+\bar{\sigma}_p^T\right]$.

2.5 Identification of internally and externally driven gene expression dynamic patterns

Because genes and meta-genes have linear relationships in terms of their expression levels as Equation (2), the components of gene expression levels in X driven by internal regulations, X_t^I can be also expressed as linear combinations of M_1 iPDPs:

$$X_t^I = W_X \tilde{X}_t^I = \sum_{p=1}^{M_1} \lambda_p^t \underbrace{W_X \tilde{V}_p^A}_{C_p^A} = \sum_{p=1}^{M_1} \lambda_p^t C_p^A \qquad (10)$$

, where $C_p^A = W_X \tilde{V}_p^A \in \aleph^{M_1 \times 1}$ includes the gene coefficients for p^{th} iPDP. The gene expression components driven by external regulations from U can be also expressed as linear combinations of M_2 ePDPs:

$$X_t^E = W_X \tilde{X}_t^E = \sum_{q=1}^{M_2} \sigma_p^t \underbrace{W_X \tilde{V}_q^B}_{C_q^B} = \sum_{p=1}^{M_2} \sigma_p^t C_q^B \qquad (11)$$

, where $C_q^B = W_X \tilde{V}_q^B \in \aleph^{M_2 \times 1}$ includes the gene coefficients for q^{th} ePDP.

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

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{B}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{X}_t^E = \sum_{q=1}^{M_2} \sigma_q^t \tilde{V}_q^B$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\sigma_q(\tilde{V}_q^B)$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $\tilde{B}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $C_p^A = W_X \tilde{V}_p^A \in \aleph^{M_1 \times 1}$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $X_t^E = W_X \tilde{X}_t^E = \sum_{q=1}^{M_2} \sigma_p^t \underbrace{W_X \tilde{V}_q^B}_{C_q^B} =$

Daifeng Wang 1/11/15 5:46 PM
Deleted: $C_q^B = W_X \tilde{V}_q^B \in \aleph^{M_2 \times 1}$

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 similar/different dynamic patterns driven by internal regulations captured by the eigenvalues of the effective system matrices between 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 Group *X* for their expression dynamics during embryonic development (Gerstein, et al., 2014). We refer to species-specific TFs as external regulations; i.e., Group *U*. 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. We focus on comparing internal dynamic patterns along with orthologous gene coefficients between worm and fly. The datasets are summarized as follows (Table 1).

We define Group *X* as 3,153 one-to-one orthologous genes between worm and fly during embryonic development, and Group *U* as all species-specific TFs (509 worm-specific TFs, 442 fly-specific TFs) (Reece-Hoyes, et al., 2005; Shazman, et al., 2014). We used their temporal gene expression levels (as measured by the RPKM values in RNA-seq) during embryonic development from the modENCODE consortium (Gerstein, et al., 2014). 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 Group *X* and five meta-genes of Group *U* 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 RESULTS

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Gerstein, et al., 2014)

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Reece-Hoyes, et al., 2005; Shazman, et al., 2014)

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Gerstein, et al., 2014)

Gene expression data during embryogenesis provide information about dynamics of genomic functions throughout developmental processes, from the conserved functions such as DNA replication to the species-specific functions such as body segmentations, but hardly reveal evolutionary GRNs that drive those developmental functions (Peter and Davidson, 2011). Thus, in order to understand the relationships between evolutionary GRNs and their driving genomic functions, we need advanced computational approaches to identify the effects of different evolutionary GRNs from gene expression data. In this study, we apply DREISS to worm and fly gene expression datasets during embryogenesis in modENCODE and identify various developmental genomic functions of worm-fly orthologous gene pairs driven by two different evolutionary GRNs, conserved (worm-fly TFs) and non-conserved (worm/fly TFs). As model organisms for developmental biology, both worm and fly have been used to study embryogenesis for decades. We found that the conserved GRNs drive similar genomic functions, but non-conserved GRNs 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 in evolution.

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Peter and Davidson, 2011)

3.1 Metagenes of orthologous genes between worm and fly have similar internal but different external principal dynamic patterns during embryonic development

We find that the meta-gene principal dynamic patterns 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) a oscillating pattern (iPDP with the complex eigenvalue) (Figure 5). In contrast to the iPDP similarities, we find that worm and fly have very different external principal dynamic patterns (ePDPs) (Figure S1). The meta-gene dynamic patterns driven by the worm-specific regulatory network; i.e., worm ePDPs consist of a varied pattern at late embryonic development (ePDP with real eigenvalue No. 1), a varied pattern at early embryonic development (ePDP with real eigenvalue No. 2), a fast increasing and then unvarying pattern (ePDP with real eigenvalue No. 3), a decaying pattern (ePDP with real eigenvalue No. 4), and an increasing pattern at late embyrnoic development (ePDP with real eigenvalue No. 5). The fly ePDPs, however, have two fast decaying patterns at early embryonic development (ePDPs with real eigenvalue No. 1 and 2), a fast increasing pattern at late embryonic develop-

ment (ePDP with real eigenvalue No. 3), and a highly increasing oscillation pattern (ePDP with complex eigenvalue). Moreover, to see the eigenvalue variations across orthologous genes, we left one gene out, and then calculated eigenvalues, which gave the eigenvalue variations shown as error bars in Figures 5 and S1. The iPDP eigenvalues vary less than ePDP in both worm and fly.

Therefore, 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 orthologous meta-genes for their different ePDPs.

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

In both worm and fly, we obtain the similar four types of internal principal dynamic patterns (iPDPs), so 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 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 6): r=0.33 (p<2.2e-16) for the highly varied pattern at late embryonic development, r=0.66 (p<2.2e-16) for the fast decaying pattern at early embryonic development, r=0.67 (p<2.2e-16) for the slowly increasing pattern during embryonic development, and r=0.73 (p<2.2e-16) for the oscillation pattern during embryonic development. This implies that, not only do the orthologous meta-genes have similar internal (conserved) regulatory effects (i.e., similar iPDPs), but the orthologous genes also have similar internally-driven expression dynamics between worm and fly because they have significantly correlated coefficients for iPDPs. The ePDPs between worm and fly generally do not have similar matches, but if we flip worm ePDP No. 3, and compare with fly ePDPs No. 4 and No. 5, they are roughly representing the fast decaying patterns.We found that the orthologous gene coefficient correlations between those ePDP patterns are much lower (r=0.12 for worm ePDP No.3 vs. fly ePDP No. 4, and r=0.18 for worm ePDP No. 3 vs. fly ePDP No. 5).

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

The ribosome produces proteins, which is an ancient process and conserved across the worm and fly, which diverged roughly a billion years ago. 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 ~200 ribosomal genes based on the GO annotations. We compared the iPDP and ePDP coefficients of ribosomal genes, and found that the iPDP coefficients are significantly larger than ePDP ones in both worm (KS-test p<0.001) and fly (KS-test p<2.2e-16) as shown in Figure 7A. This means that the ribosomal gene expression is significantly more driven by the conserved regulatory network than by the species-specific regulatory network, which is consistent with ribosomal genes having conserved functions during embryonic development.

Daifeng Wang 1/11/15 5:46 PM
Deleted: involved in ribosome biogenesis

The orthologous genes related to signal transduction for cell-cell communication, which is a significantly more recent evolutionary adaptation relative to ribosomes, however, exhibit the opposite trend. We found that 320 signaling genes from GO annotations have significantly larger ePDP coefficients than iPDP ones in both worm (KS-test p<7e-4) and fly (KS-test p<6e-4), as shown in Figure 7B. 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 typically being associated with species-specific functions, such as body plan establishment and cell differentiation.

3.4 DNA replication and Proteasome machinery are enriched in orthologous genes with high coefficients for principal dynamic patterns where expression levels increase fast at early embryonic development and remain flat afterwards (fast-growing patterns)

We next turn to the biological meaning of individual iPDPs and ePDPs. For the fast-decaying pattern (2nd iPDP), we found that the DNA replication is significantly enriched in Top 300 (~10%) orthologous genes that have 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 in Figure 8). 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 (Bate and Martinez Arias, 1993; Baugh, et al., 2003). The mRNA

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Bate and Martinez Arias, 1993; Baugh, et al., 2003)

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 8), 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. Also, we did not find any enrichment of DNA replication in top genes of other iPDP and ePDP patterns. Therefore, the iPDP patterns identified by our method reveal basic 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 TFs, we also identified a fast growing pattern driven by non-conserved TFs for those two species. The Top 300 orthologous genes (~10%) with fast-growing worm ePDP and fly ePDP (i.e., driven by species-specific regulatory networks) shared 36 orthologous genes. 10 of them encode genes in the proteasome complex (p-value<1.2e-9). Protein degradation is not only a key process in apoptosis, but also throughout the whole process of development (DeRenzo and Seydoux, 2004). 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 (Shen-Orr, et al., 2010). In our study, after separating the effect of conserved and non-conserved regulatory networks, the protein degradation is significantly enriched in the genes majorly driven by species-specific TFs.

Daifeng Wang 1/11/15 5:46 PM
Deleted: (DeRenzo and Seydoux, 2004)

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Shen-Orr, et al., 2010)

Besides the 36 shared genes in the fast-growing pattern driven by species-specific TFs, there are additional observations that we find interesting. Among the Top 300 worm orthologous genes with fast-growing ePDPs, genes involved in calcium ion binding (p-value<2e-6), GTP binding (p-value<7e-3) and neuron differentiation (p-value<0.05) are over-represented, which implies that they are activated in the early stage of embryogenesis by worm-specific TFs. This observation indicates the GRN of these genes have evolved after the speciation. Proteins involved in calcium ion binding or GTP binding usually play a role in cell signal transduction (Aspenstrom, 2004). In fact, the genes involved in Wnt signaling and MAPK signaling both exhibit a two-fold change.

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Aspenstrom, 2004)

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-value<5e-10). It is well-known the Wnt signaling in worms starts as early as at the 4-cell stage, when one cell receives the signal and starts differentiation (Sawa and Korswagen, 2013). The separation of regulatory effects showed that the expression of genes involved in signaling is more controlled by the species-specific TFs. The enrichment of energy generation in the Top 300 fly genes with a fast-growing ePDP is probably indicative of the large energy requirement during fly embryogenesis (Tennessen, et al., 2014), 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. In addition, the lack of signaling enrichment might be due to different sampling time points. Since each of the first 10 cell cycles takes less than 10 minutes in the fly embryo (Gilbert, 2000), the 2 hour time interval may not have the resolution to capture the early regulatory events.

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Sawa and Korswagen, 2013)

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Tennessen, et al., 2014)

Daifeng Wang 1/11/15 5:46 PM
Deleted: (Gilbert, 2000)

4 DISCUSSION

In this paper, we developed 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 coming outside the group (external regulatory subsystem). We applied DREISS to the time-series gene expression datasets for worm and fly embryonic developments from the modENCODE project (Gerstein, et al., 2014), and compared the worm-fly orthologous gene expression dynamic patterns driven by the conserved regulatory network (i.e., regulation effects from orthologous TFs), and the patterns driven by the species-specific regulatory networks (i.e., regulation effects from worm or fly specific TFs). This work can be 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.

Daifeng Wang 1/11/15 5:46 PM
Deleted:).

Daifeng Wang 1/11/15 5:46 PM
Deleted: embryogenesis

Daifeng Wang 1/11/15 5:46 PM
Deleted: of

Daifeng Wang 1/11/15 5:46 PM
Deleted: in

Daifeng Wang 1/11/15 5:46 PM
Deleted: ,

We emphasize that DREISS is a general-purpose method (a free downloadable tool at github.com/gersteinlab/dreiss). Users can define the internal group (*X*) and external group (*U*) 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. Also, 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 expression patterns in development. The expression patterns driven by human-specific regulatory factors potentially help us understand human-specific developmental processes along with associated human genes.

Daifeng Wang 1/11/15 5:46 PM
Deleted: .

Daifeng Wang 1/11/15 5:46 PM
Deleted: then comparing

Daifeng Wang 1/11/15 5:46 PM
Deleted: , leading to the discovery of novel

Daifeng Wang 1/11/15 5:46 PM
Deleted: genes specifically

Daifeng Wang 1/11/15 5:46 PM
Deleted: with our development

Daifeng Wang 1/11/15 5:46 PM
Deleted: ACKNOWLEDGMENTS

REFERENCES

Aspenstrom, P. (2004) Integration of signalling pathways regulated by small GTPases and calcium, *Biochimica et biophysica acta*, **1742**, 51-58.

Bansal, M., Della Gatta, G. and di Bernardo, D. (2006) Inference of gene regulatory networks and compound mode of action from time course gene expression profiles, *Bioinformatics*, **22**, 815-822.

Bate, M. and Martinez Arias, A. (1993) *The Development of Drosophila melanogaster*. Cold Spring Harbor Laboratory Press, Plainview, N.Y.

Baugh, L.R., *et al.* (2003) Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome, *Development*, **130**, 889-900.

Brogan, W.L. (1991) *Modern control theory*. Prentice Hall, Englewood Cliffs, N.J.

Chu, S., *et al.* (1998) The transcriptional program of sporulation in budding yeast, *Science*, **282**, 699-705.

DeRenzo, C. and Seydoux, G. (2004) A clean start: degradation of maternal proteins at the oocyte-to-embryo transition, *Trends in cell biology*, **14**, 420-426.

Gerstein, M.B., *et al.* (2014) Comparative analysis of the transcriptome across distant species, *Nature*, **512**, 445-448.

Gilbert, S.F. (2000) *Developmental biology*. Sinauer Associates, Sunderland, Mass.

Huang, S. and Ingber, D.E. (2006) A non-genetic basis for cancer progression and metastasis: self-organizing attractors in cell regulatory networks, *Breast disease*, **26**, 27-54.

Irie, N. and Kuratani, S. (2011) Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis, *Nature communications*, **2**, 248.

Kalinka, A.T., *et al.* (2010) Gene expression divergence recapitulates the developmental hourglass model, *Nature*, **468**, 811-814.

Kim, P.M. and Tidor, B. (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data, *Genome research*, **13**, 1706-1718.

Levin, M., *et al.* (2012) Developmental milestones punctuate gene expression in the Caenorhabditis embryo, *Developmental cell*, **22**, 1101-1108.

Peter, I.S. and Davidson, E.H. (2011) Evolution of gene regulatory networks controlling body plan development, *Cell*, **144**, 970-985.

Rangel, C., *et al.* (2004) Modeling T-cell activation using gene expression profiling and state-space models, *Bioinformatics*, **20**, 1361-1372.

Reece-Hoyes, J.S., *et al.* (2005) A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks, *Genome biology*, **6**, R110.

Saeys, Y., Inza, I. and Larranaga, P. (2007) A review of feature selection techniques in bioinformatics, *Bioinformatics*, **23**, 2507-2517.

Sawa, H. and Korswagen, H.C. (2013) Wnt signaling in C. elegans, *WormBook : the online review of C. elegans biology*, 1-30.

Shazman, S., *et al.* (2014) OnTheFly: a database of Drosophila melanogaster transcription factors and their binding sites, *Nucleic acids research*, **42**, D167-171.

Shen-Orr, S.S., Pilpel, Y. and Hunter, C.P. (2010) Composition and regulation of maternal and zygotic transcriptomes reflects species-specific reproductive mode, *Genome biology*, **11**, R58.

Tennessen, J.M., *et al.* (2014) Coordinated metabolic transitions during Drosophila embryogenesis and the onset of aerobic glycolysis, *G3*, **4**, 839-850.

Vilar, J.M. (2006) Modularizing gene regulation, *Molecular systems biology*, **2**, 2006 0016.

Wang, D., *et al.* (2012) Principal-oscillation-pattern analysis of gene expression, *PloS one*, **7**, e28805.

Wang, D.*, et al.* (2012) Eigen-genomic system dynamic-pattern analysis (ESDA): modeling mRNA degradation and self-regulation, *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, **9**, 430-437.
Yanai, I.*, et al.* (2011) Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility, *Developmental cell*, **20**, 483-496.