Journal of Physics: Conference Series **490** (2014) 012145 doi:10.1088/1742-6596/490/1/012145

Network inference of AP pattern formation system in *D.melanogaster* by structural equation modeling

S Aburatani and H Toh

Computational Biology Research Center, AIST, AIST Tokyo Waterfront Bio-IT Research Building, 2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan

E-mail: s.aburatani@aist.go.jp

Abstract. Within the field of systems biology, revealing the control systems functioning during embryogenesis is an important task. To clarify the mechanisms controlling sequential events, the relationships between various factors and the expression of specific genes should be determined. In this study, we applied a method based on Structural Equation Modeling (SEM), combined with factor analysis. SEM can include the latent variables within the constructed model and infer the relationships among the latent and observed variables, as a network model. We improved a method for the construction of initial models for the SEM calculation, and applied our approach to estimate the regulatory network for Antero-Posterior (AP) pattern formation in *D. melanogaster* embryogenesis. In this new approach, we combined cross-correlation and partial correlation to summarize the temporal information and to extract the direct interactions from the gene expression profiles. In the inferred model, 18 transcription factor genes were regulated by not only the expression of other genes, but also the estimated factors. Since each factor regulated the same type of genes, these factors were considered to be involved in maternal effects or spatial morphogen distributions. The interpretation of the inferred network model allowed us to reveal the regulatory mechanism for the patterning along the head to tail axis in *D. melanogaster*.

1. Introduction

Clarifying the process that generates an organism's body plan is one of the fascinating themes, and the process of *D. melanogaster* embryo formation is a suitable model for studying embryogenesis. During *D. melanogaster* embryogenesis, anterior-posterior (AP) patterning is generated by serial transcriptional regulation, beginning with maternally supplied transcripts, through gap genes and pair-rule genes, and finally segment polarity genes [1]. The expression of these genes generates gradients of their proteins in the embryo, and thus these gradients determine the expression of the following genes [2]. The regulation of developmental genes has been extensively studied, but many aspects remain unexplained and we are far from a complete understanding of pattern formation.

To obtain better insights into the transcriptional regulatory mechanism for pattern formation, a gene regulatory network is useful [3, 4]. Since the underlying mechanism of transcriptional regulation is the localized expression of transcription factor genes at specific times and places, revealing the regulatory networks between these genes would provide a schema of developmental regulation in embryogenesis. Furthermore, the influences from several types of cellular components should be considered as regulatory factors of gene expression. Thus, we have developed a new approach, based on Structural Equation Modeling (SEM), to investigate the regulatory relationships between different cellular components [5, 6].

In this study, we developed a new method to assume an initial model for SEM calculation, and improved our SEM approach for inferring complicated transcriptional regulation in *D. melanogaster* embryogenesis. We applied our approach to 18 transcription factor genes with regulated expression for AP pattern formation. Using our methods, the regulatory factors for AP pattern formation were estimated by factor analysis, and the regulatory relationships from the regulators to the genes were estimated by SEM.

2. Materials & Methods

2.1. Expression data and selected genes

For the construction of the initial model, we utilized the expression profiles of 12,868 genes measured during 28 time points, covering the entire 24-h period during *D. melanogaster* embryo development [7]. All expression data in *D. melanogaster* embryo cells were downloaded from GEO Database (http://www.ncbi.nlm.nih.gov/geo/). Among the empirically identified transcription factor genes for AP pattern formation [8], 18 genes displayed sufficient expression profiles, and these were utilized to infer the regulatory network.

2.2. Construction of the initial model

To construct an initial model, we developed a new method that combines cross correlation and partial correlation. Causal relationships between genes were detected by cross correlation, given by

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\left\{\sum_{i=1}^{N}(x_i - \bar{x})^2\right\}^{1/2}\left\{\sum_{i=1}^{N}(y_{i+d} - \bar{y})^2\right\}^{1/2}} \tag{1}$$

where x_i and y_i are two time series data with N time points, and d is the time-lag between variables x and y. In this case, each gene pair was calculated with d = -5, … , +5. The eleven absolute values of cross correlations were compared, and the highest absolute value was arranged in a fundamental cross correlation matrix. The corresponding time-lag value d was arranged as a matrix element in a discrete matrix.

The direct interactions were estimated by the partial correlation coefficients calculated from the cross correlation matrix, as follows:

$$r_{ij|rest} = \frac{r^{ij}}{(r^{ii})^{1/2}(r^{jj})^{1/2}} \tag{2}$$

where $r_{ij|rest}$ is the partial correlation coefficient between variables i and j, given the rest, and r^{ij} is the (i, j) element in the reverse of the cross correlation coefficient matrix. To detect the significant value of the partial correlation coefficient, we use the statistic

$$t_{ij} = \frac{r_{ij|rest}(n-q-2)^{1/2}}{(1-r_{ij|rest}^2)^{1/2}} \tag{3}$$

where q is the number of fixed variables and n is the number of samples. The statistic t_{ij} is distributed according to the t distribution. Thus, the relationship between the expression profiles can be tested by the t-test. In this study, we selected the significant gene pairs at $p < 0.01$.

Finally, we combine partial correlation and cross correlation. The regulatory directions, which were estimated by the sign of the time-lag values, were added to the extracted gene pairs by PCC. The procedure developed for constructing the initial model is displayed in Figure 1.

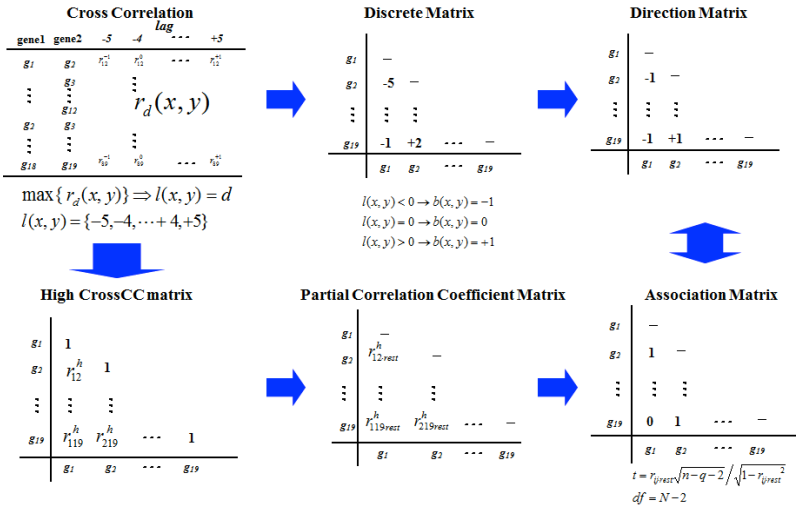

Figure 1. Procedure for initial model construction.

2.3. Factor Analysis

To determine the optimal number of factors for inclusion in the network model as latent variables, we performed a factor analysis. In the factor analysis, the covariance matrix between the observed variables Σ is structurized by parameters, as follows:

$$\Sigma = Var[X] = \Lambda \Phi \Lambda^{t} + \Psi^{2} \tag{4}$$

where Ψ^2 is the covariance matrix of error terms, Λ is the factor loading matrix of latent variables, and Φ is the covariance matrix among factors. From this structurized matrix, the values of the matrix and the variances of the error terms are estimated. In this study, the Kaiser criterion states and the scree plot were utilized to estimate the number of factors. The number of latent variables was suggested by a principal factor method with promax rotation, which is a general method for rotating factors to fit a hypothesized structure of latent variables.

2.4. Structural Equation Modeling (SEM)

In this study, the regulatory model is defined as follows:

$$\begin{bmatrix} f \\ v \end{bmatrix} = \begin{bmatrix} \mathrm{B} & \Gamma \\ \Theta & \mathrm{K} \end{bmatrix} \begin{bmatrix} f \\ v \end{bmatrix} + \begin{bmatrix} \delta \\ \varepsilon \end{bmatrix} \tag{5}$$

Here, f is a vector of p latent variables, and v is a vector of q observed variables. The matrix B is a $p \times p$ matrix representing the relationships between latent variables; Γ is a $p \times q$ matrix representing the causal relationships between observed and latent variables; Θ is a $q \times p$ matrix representing the effectiveness of latent variables to the observed variable; and K is a $q \times q$ matrix representing the regulatory relationships between observed variables. In the SEM analysis, the parameter estimation was performed by comparing the actual covariance matrix S, calculated from the measured data, with the estimated covariance matrix $\Sigma(\theta)$ of the constructed model. The maximum likelihood method was utilized as a fitting function to estimate the model parameters. To optimize the network model, we developed an iteration algorithm by using fitting scores and modification index (MI) scores, which measure how much the chi-square statistic is expected to decrease if a particular parameter setting is constrained. The SEM software package SPSS AMOS 17.0 (IBM, USA) was used.

3. Results & Discussion

By our combination of cross correlation and partial correlation, we estimated 35 causal relationships between 18 genes. Furthermore, the application of factor analysis provided information about

unobserved effective factors, and 4 factors were estimated as regulators for the 18 genes. Thus, the optimal model was inferred from an initial model with the 18 genes and the detected factors. The whole network is displayed in Figure 2a. In Figure 2b, the order of the 18 genes is consistent with the biological knowledge of them, even though the order was disturbed in the initial model. The inclusion of latent variables into an inferred model is considered to be suitable for representing a biological phenomenon. Furthermore, Figure 2c shows that the known temporal and spatial gene expressions during *D. melanogaster* development are well reflected by the inferred networks between genes.

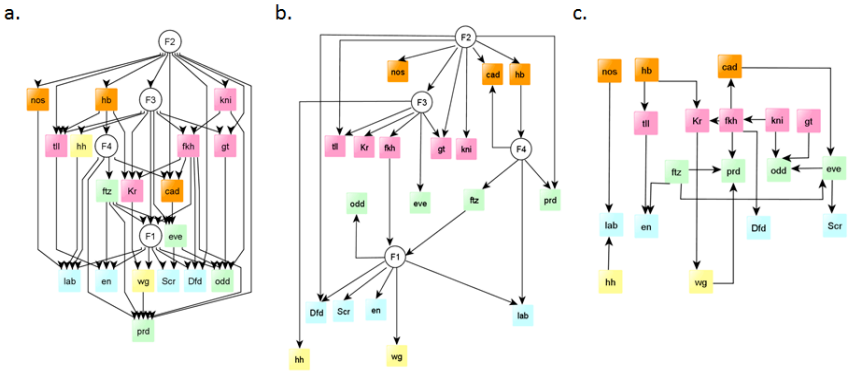

Figure 2. Inferred networks. (a) Whole network model. (b) Relationships between factors and genes. (c) Relationships between genes. Orange: Maternal, Pink; Gap, Green: Pair-rule, Blue; Segment, Yellow; Signal.

Since F2 was a positive regulator of all maternal genes and some gap genes, F2 was considered as Bicoid protein density in the early embryo. Furthermore, F3 regulated many gap genes positively, but negatively regulated one pair-rule gene, therefore F3 was considered to be an inducer of the gap genes' process by repressing pair-rule gene expression.

4. Conclusion

In this study, the spatial and temporal controls that function in the developmental process were identified by our inferred network. In the inferred network, not only the effects of transcription factor proteins, but also the protein densities were suggested as latent variables. The network inference by SEM is applicable to clarify the control of gene expression by intracellular factors.

References

[1] Pankratz M J and Jäckle H 1993 *Blastoderm segmentation. In The Development of Drosophila melanogaster*. (Cold Spring Harbor: Cold Spring Harbor Laboratory Press) 467-516

[2] Rivera-Pomar R and Jäckle H 1996 From gradients to stripes in Drosophila embryogenesis: Filling in the gaps *Trends Genet*. **12** 478–483

[3] Friedman N, Linial M, Nachman I and Pe'er D 2000 Using Bayesian networks to analyze expression data *J. Comput. Biol*. **7** 601-620

[4] Akutsu T, Miyano S and Kuhara S 2000 Inferring qualitative relations in genetic networks and metabolic pathways *J. Comput. Biol*. **7** 331-343

[5] Aburatani S 2011 Application of Structure Equation Modeling for inferring a serial transcriptional regulation in yeas *Gene. Regul. Syst. Biol*. **5** 75–588

[6] Aburatani S 2011 Network Inference of pal-1 Lineage-Specific Regulation in the C. elegans Embryo by Structural Equation Modeling *Bioinformation* **8** 652–657

[7] Hooper S D, Boué S, Krause R, Jensen L J, Mason C E, Ghanim M, White K P, Furlong E E and Bork P 2007 Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis *Mol. Sys. Biol*. **3** 1-11

[8] MacArthur S *et al*. B 2009 Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions *Genome Biol*. **10** R80