BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(14)

Network inference of *pal-1* lineage-specific regulation in the *C. elegans* embryo by structural equation modeling

Sachiyo Aburatani

Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-4-7 Aomi, Koto-ku, Tokyo, Japan; Sachiyo Aburatani – Email: s.aburatani@aist.go.jp; Phone: +81 3 3599 8712

Received July 06, 2012; Accepted July 07, 2012; Published July 21, 2012

Abstract:
The elucidation of spatial and temporal control during developmental stages is one of the central tasks for systems biology, and a variety of intracellular factors are known as regulators for specific gene expression. The activity information of those various factors is not directly reflected in their gene expression profiles. Hence, a method based on Structural Equation Modeling (SEM) is described. SEM can include the latent variables within the constructed model and infer the relationships among latent and observed variables, as a network model. An improved SEM approach for the construction of an optimal model is applied to infer the regulatory network for the determination of C lineage fate in *C. elegans* development. The inferred network model shows that the 13 analysed transcription factor genes were regulated by several other factors in addition to *pal-1* expression. The other regulatory factors are those involved in protein accumulation and localization as important regulatory factors for normal development. Those regulatory factors were regulated sequentially in the network model. The regulation of the known *pal-1* regulated genes was dependent on this sequential control of the regulatory factors. The interpretation of the network model shows insights to the complex regulation occurring during the C lineage determination by *pal-1.*

Keywords: Structural Equation Modeling, Embryo, Gene Network, Developmental Control

Background:
In multi-cellular organisms, asymmetric cell division and cell differentiation are essential for normal development. Asymmetric cell division in embryogenesis occurs to generate body axes, and cell differentiation determines cell fate **[1-3]**. Through these crucial developmental periods, a cell becomes specialized to construct tissues and organs, according to its fate **[4]**. These developmental events are controlled to divide the developmental determinants into suitable descendants **[5, 6]**, but much remains to be elucidated about the regulatory system in the early embryo. Since embryogenesis is controlled spatially and temporally, the entire regulatory system in early embryonic development is incredibly complicated **[7]**.

To gain a better understanding of the role of developmental control, a gene regulatory network is useful. The application of various algorithms, including Boolean and Bayesian networks and graphical Gaussian modeling (GGM), to gene expression profiles allows us to infer complex functional gene networks **[8-11]**. One of the clues toward revealing the developmental regulation in the early embryo is to clarify the factors influencing cell fate determination during embryogenesis. Cell fate is usually determined in normal development, as the course from the zygote to the complete organism. At the early stage of development, cell fate determination is executed by the regulated translation of stored maternal mRNAs and the accommodation of protein activity **[12, 13]**. Furthermore, gene expression control by transcription factors is also needed for cell

fate determination and cellular differentiation **[14]**. Corresponding to the gene expression changes, cell fate determination and cellular differentiation are activated during the developmental stage. This means that there are several types of cellular factors that regulate cell differentiation in embryogenesis. Therefore, to reveal the regulatory networks in embryogenesis, such as how a cell's fate is determined, a new network inference approach is needed.

Recently, I developed a new statistical approach, based on Structural Equation Modeling (SEM) in combination with factor analysis and a four-step procedure **[15]**. I developed this approach to reveal a serial transcriptional regulation system mediated by transcription factor proteins, by using information from only gene expression profiles, and no protein information. One of the significant features of SEM is the inclusion of latent variables into the constructed model, which allows the inferred model to include transcription factor proteins as latent variables, and genes as observed variables. This method estimates the significant interactions between variables. In the constructed model, linear relationships among variables are assumed to minimize the differences between the fitted model covariance matrix and the calculated sample covariance matrix. This approach allowed me to reconstruct the hierarchical model of transcriptional regulation that involves different cellular components, proteins and DNA.

The clarification of cell fate determinants and their effects by my SEM approach is considered to be useful for revealing the developmental control occurring in the *C. elegans* early embryo. In *C. elegans*, cell division during embryonic development, from the zygote to all 959 somatic cells, can be traced **[16].** Through the early stages in embryogenesis, 5 founder cells, AB, MS, E, C and D, are produced by asymmetric cell division to generate distinct sets of somatic cells **[16].** Among these founder cells, the C blastomere mainly gives rise to muscle and epidermis, and the cell fate of the C blastomere is regulated through a genetically defined transcriptional cascade of activation by the protein PAL-1 **[17]**. Based on previous investigations, PAL-1 is considered to maintain the identity of the C blastomere at the eight-cell stage in embryogenesis. The translation of the maternal *pal-1* mRNA is known to be sequentially restricted until the four-cell stage in embryogenesis, and the C blastomere fails to develop in the absence of maternal PAL-1 activity **[7]**. Furthermore, ectopic PAL-1 activity gives rise to muscle and epidermal cells by the C-like lineage in the other somatic lineages, in the absence of maternal PAL-1 **[18]**.

Here, I applied the SEM approach to reveal the *pal-1*-mediated regulation in embryogenesis, by using the expression profiles of *pal-1*-dependent genes, which have been measured to clarify the *pal-1* effect. The PAL-1 transcription factor protein is considered to regulate 12 other transcription factor genes, including uncharacterized proteins, and those PAL-1 target genes have been experimentally confirmed to affect the C-lineage differentiation **[17, 19, 20]**. Even though some of the regulatory pathways from *pal-1* to its target genes have been identified, the functional mechanisms of the *pal-1* mRNA or PAL-1 protein remain unclear. In this study, I employed an improved SEM approach to extract the factors for cell fate determination and to reconstruct a regulatory network model among the *pal-1* target genes. Using this method, the determinants of cell fates were extracted by a factor analysis. I could estimate not only the unobserved regulators for gene expression, but also the significant regulation pathways, from regulators to genes. The resulting gene expression profiles revealed the well coordinated developmental control by *pal-1*.

Methodology:

Expression data

I combined two early embryonic expression profiles in *C. elegans* for the SEM calculation. One profile is GSE2180, including 123 samples measured by Baugh *et al.* **[17],** and the other profile is GSE9665, including 74 samples measured by Yanai *et al.* **[19].** In both experiments, 22,625 gene expression profiles were measured to reveal the C-lineage-specific genes. Among them, the following 12 genes were identified as transcription factors that are regulated by *pal-1* in the C-lineage embryo: *tbx-8*, *tbx-9*, *elt-1*, *hnd-1*, *scrt-1*, *lin-26*, *nhr-25*, *vab-7*, *elt-3*, *hlh-1*, *unc-120* and *nob-1*. Furthermore, *pop-1* is considered to be associated with cell fate decision at the four-cell stage. Thus, I analyzed the expression profiles of 14 genes, including *pal-1*, 12 *pal-1*-regulated genes and *pop-1*, which are considered to function in the C-lineage embryo.

Factor Analysis

The network analysis by SEM includes two steps: parameter fitting and model structure fitting. To assume the model structure, I selected the optimal number of factors for inclusion in the network model as latent variables, by performing a factor analysis. In the factor analysis, the covariance matrix between the observed variables $\sum$ is structurized by parameters, as follows:

$$\sum = Var[X] = \Lambda \Phi \Lambda^t + \Psi^2 \rightarrow \textbf{(1)}$$

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 matrix Λ and the variances of the error terms are estimated. In this study, the Kaiser criterion states and the scree plot were utilized to estimate a number of factors. In the Kaiser criterion, the number of factors is equal to the number of eigen values of the covariance matrix that are greater than one. The number of latent variables was suggested by a principal factor method with varimax rotation, which is a general method for rotating factors to fit a hypothesized structure of latent variables.

Structural Equation Modeling (SEM)

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

$$y = \Lambda' n + \Gamma y + \varsigma \rightarrow \textbf{(2)}$$

Here, y is a vector of p observed variables (genes), and n is a vector of q latent variables (regulatory factors). The effectiveness of the factors to the genes is represented by Λ', and the relationships between the genes are represented by Γ, as matrix forms. Errors that affect genes are denoted by ς. According to this model definition, the model covariance matrix $\Sigma(\theta)$ is represented by parameters. In the SEM analysis, the parameter estimation was performed by comparing the actual covariance matrix Σ, calculated from the measured data, with the estimated covariance matrix $\Sigma(\theta)$ of the constructed model.

I used the maximum likelihood method as a fitting function to estimate the model parameters. The SEM software package SPSS AMOS 17.0 (IBM, USA) was used to fit the model to the data.

Iterations for the Optimal Model

The constructed models are evaluated by their structures, in comparison to the measured data. To detect the quantitative similarity between a constructed model and an actual relationship, fitting scores are usually utilized. By using these scores, I developed an iteration algorithm to optimize the model, as follows:

Step 1: Reconstruction of the network model without a non-significant edge; Step 2: Re-calculation of all parameters from the reconstructed model; Step 3: Iteration of Steps 1 and 2 until all edges become significant; Step 4: Addition of a possible causal edge to the reconstructed model by the Modification Index (MI); Step 5: Iteration from Steps 1 to 3 to confirm that the other edges in the model are significant; Step 6: Determination of significant relationships among error terms.

The MI measures how much the chi-square statistic is expected to decrease if a particular parameter setting is constrained. After all of the edges are significant and all of the MI scores are lower than 10.0 in the constructed model, the significant relationships between the error terms are estimated by the MI scores. The relationships among the error terms have no direction, and thus they are a correlation between error terms. These relationships were used for the calculations, but were not incorporated into the network.

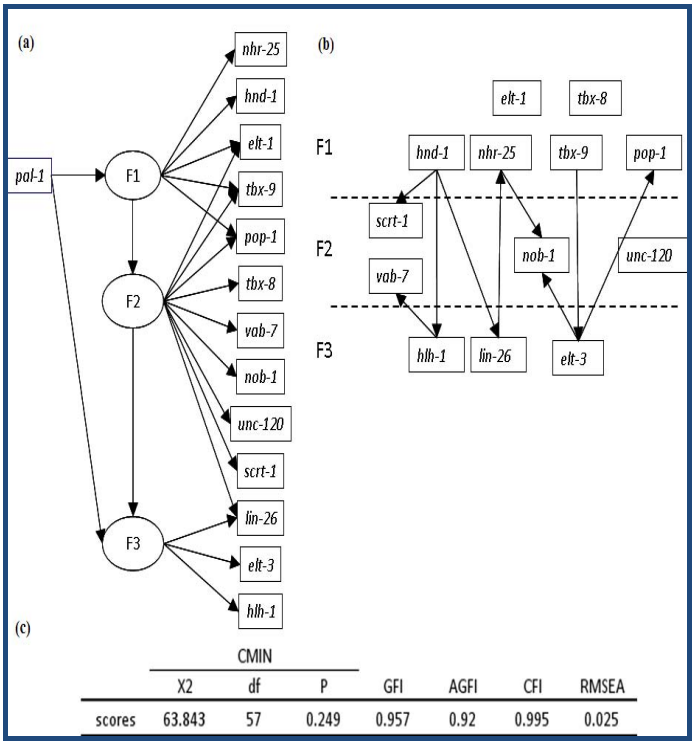

Figure 1: Inferred network model of *pal-1* regulation. The estimated network structure of the *pal-1* regulatory system shows for lineage-specific differentiation. Genes, which are observed variables, are displayed as rectangles, and estimated regulatory factors, which are latent variables, are displayed as circles. Arrows show the causal relationships among the variables in the model. Error terms are omitted in this figure, but all error terms were calculated by SEM. The relationships between the errors are considered to represent other regulatory systems in the cell. For simplicity, these relationships are not shown. **(a)** Network model between genes and regulatory factors; **(b)** Relationships between *pal-1*-dependent genes. Each gene is classified by its regulatory factor, shown on the left side; **(c)** Goodness-of-fit scores. The calculations for these scores included the relationships between errors. Four criteria were mainly used: GFI>0.90, AGFI>0.90, CFI >0.90 and RMSEA<0.05. All four scores indicated that the model fit the measured data well.

Discussion:

Regulatory factors of pal-1 regulated genes

To reveal the regulatory network in early embryogenesis, I first detected the intracellular regulatory factors for gene expression. In this study, 14 genes are described as observed variables, and the regulatory factors are arranged as latent variables in the network model. I utilized factor analysis to reveal the underlying structure among the variables. First, exploratory factor analysis (EFA) was applied to detect the number of regulatory factors for the expression of the 14 genes, since EFA is commonly used for identifying the set of latent variables with effects on the observed variables.

From the compiled expression profiles of the 14 genes measured under 197 conditions, 4 regulatory factors were extracted by the first EFA. To clarify the percent of variance in each gene explained by the extracted factors, the communality of each gene was calculated from the sum of the squared factor loading for all factors. According to the first EFA, the communality of *pal-1* was lower than 0.1, and this may be interpreted as meaning that *pal-1* expression was not affected by the regulatory factors. In this study, *pal-1* is considered as an initiator of the C-lineage, and thus independent *pal*-1 expression was reasonable.

I applied the second EFA to the expression profiles of 13 genes without *pal-1*, and 3 factors were extracted. The communality and factor loading of each factor are displayed in **Table 1(see supplementary material)**. In **Table 1,** the genes are divided into three clusters according to their factor loading: the genes mainly regulated by factor 1 (F1), the genes mainly regulated by factor 2 (F2), and the genes mainly regulated by factor 3 (F3). **Table 1** also shows the stage at which each gene was detected, which was empirically confirmed by Yanai *et al.* **[19].** The genes that had been detected as initiators of the C-lineage, *tbx-8* and *tbx-9*, were regulated by F1. Furthermore, *elt-1* and *scrt-1*, which were detected at the early stage in embryogenesis, were also regulated by the same factor. The genes that were mainly regulated by F2 were also detected at an early stage in cell division, but not as initiators. From these features, F1 and F2 may be regulators that function at an early stage of embryogenesis. A focus on the detected cell type of each gene indicated that F3 regulates all of the genes that were detected as only epidermal. Even though one muscle gene was also regulated by F3, this muscle gene was detected at a later stage in embryogenesis. Thus, the features of F3 are considered to be different from those of the other factors.

Regulatory networks for C lineage fate

Before the SEM calculation, I assumed an initial model that includes both the latent and observed variables. The restrictions of the initial model were determined as follows: 1) three latent variables were arranged as the effective regulatory factors of 13 TF genes, 2) regulatory relationships were assumed from the latent variables to the observed variables, depending on the values of the factor loadings, and 3) the observed variable *pal-1* was arranged at the starting point in the initial model, since *pal-1* is considered to be an initiator of the C-lineage fate in this study. With these restrictions, I applied the modified four-step procedure developed in my previous investigation **[15],** and an initial model was constructed with *pal-1* and the other TF genes connected by latent variables. All possible regulatory patterns between *pal-1* and the three latent variables were evaluated by SEM, and the optimal regulatory model was selected as the most suitable network shape for expression profiles.

The inferred network is shown in **(Figure 1)**. By my iteration steps developed for model optimization, all edges within the model were significant ($p<0.05$). The causalities between the factors and the genes are shown in **(Figure 1a)**, and the relationships among the genes are shown in Figure 1b. It is known that *lin-26* and *hnd-1* repress *nhr-25* and *hlh-1*, respectively **[19]**, and those known regulatory relationships were well described in **(Figure 1b)**. The regression weight between *lin-26* and *hnd-1* and that between *hnd-1* and *hlh-1* were estimated as negative values, which means repression control. To evaluate the model fitting, I utilized general goodness-of-fit scores, as follows: goodness-of-fit index (GFI), adjusted GFI (AGFI), CFI, and RMSEA. These indices have threshold values as criteria to decide whether the model is suited to the measured data, and Figure 1c shows that all of the indices indicated that the inferred model is suited to the expression data.

In **(Figure 1a)**, the three latent variables are regulated sequentially, and almost all of the early embryo genes were regulated by the first latent variable, F1. The factors were expected to be regulated by *pal-1*, such as encoding, maternal mRNA division, protein activity by accumulation, and so on. Thereby, the three factors were interpreted by regulatory orders in the resulting model. **Figure 2** shows the interpretation of the latent variables. Factor F1 was considered to be the quantity of the PAL-1 protein, since it is only regulated by the *pal-1* mRNA, and thus the regulatory relationships between *pal-1* and F1 were considered as "translation". Factor F2 was only regulated by F1, and it regulated early embryonic genes and other genes. Sequential restriction of PAL-1 activity is known to occur, and thus F2 was interpreted as the PAL-1 activity that was dependent on the blastomere. Factor F3 was regulated by *pal-1* and F2, and F3 mainly regulated epidermal genes. All of the F3 regulated genes were detected when the C blastomere divides into 31 cells, even though the other genes were detected at the former stage in cell differentiation, Thus, F3 was considered as a regulator functioning after cell division. Actually, the *pal-1* mRNA is known to be partitioned into daughter cells during cell differentiation. Thus, the causality between *pal-1* and F3 was estimated as mRNA segregation, and F3 was considered as the localization for *pal-1* spatial control.

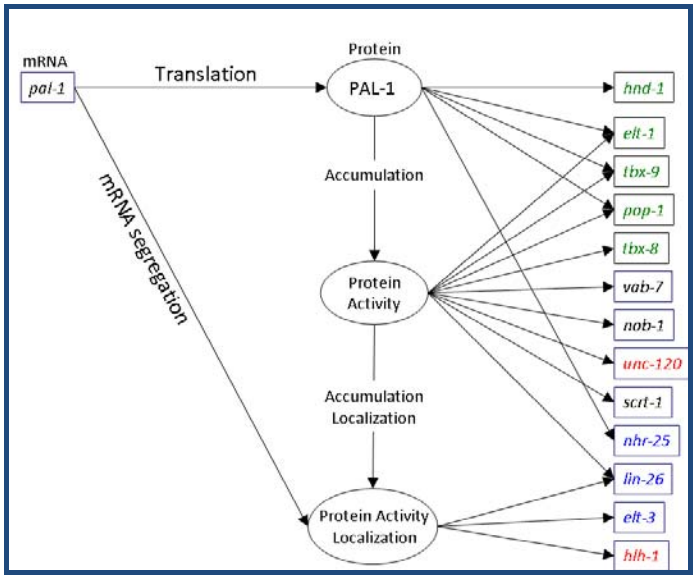

Figure 2: Biological interpretation of the inferred model. Biological interpretations of regulatory factors are expressed within the network. All genes are displayed as rectangles, and the color of the gene name indicates the detected cell type or developmental phase in embryogenesis: Green: Early embryo, Red: Muscle cell, Blue: Epidermal cell and Black: Both muscle and epidermal cells. The *pal-1* rectangle indicates the quantity of *pal-1* mRNA, and thus the first regulatory factor is considered to be the PAL-1 protein. From the sequential regulation of the regulatory factors, all factors and relationships were interpreted.

Conclusion:

In normal development, genes are regulated by many factors; however, the ambiguity of the underlying mechanisms is one of the serious obstacles to artificial cell differentiation. SEM is a powerful approach to estimate the gene regulatory network in cell differentiation. The spatial and temporal control mechanisms of *pal-1* have been solved by my inferred network, since SEM is a useful method for constructing a regulatory network including unknown factors. The inferred network shapes reflect the features of cell fate determination for the C lineage, which is regulated by *pal-1*. The effects of protein accumulation and localization were suggested as latent variables in addition to PAL-1 regulation in the inferred model. SEM will be applicable to a wide number of gene networks, to clarify the control of gene expression by intracellular factors as biological data gets accumulated. The ability to identify expression profiles and the corresponding biological functions is expected to provide applications for SEM for the inference of regulatory mechanisms in cell differentiation.

References:

[1] Neumüller RA & Knoblichet JA, *Genes Dev.* 2009 **23**: 2675 [PMID: 19952104]
[2] Kaletta T *et al. Nature*. 1997 **390**: 294 [PMID: 9384382]
[3] Lin R *et al. Cell*. 1998 **92**: 229 [PMID: 9811572]
[4] Huang S *et al. Development.* 2007 **134**: 2685 [PMID: 17567664]
[5] Wood WB *et al. J Nematol.* 1982 **14:** 267 [PMID: 19295708]
[6] Hwang SY, *BMB Rep.* 2010 **43**: 69 [PMID: 20193124]
[7] Hunter CP & Kenyon C, *Cell*. 1996 **87**: 217 [PMID: 8861906]
[8] Akutsu T *et al. J Comput Biol.* 2000 **7**: 331 [PMID: 11108466]

[9] Friedman N *et al. J Comput Biol.* 2000 **7**: 601 [PMID: 11108481]
[10] Aburatani S *et al. Nucleic Acids Res.* 2005 **33**: W659 [PMID: 15980557]
[11] Aburatani S & Horimoto K, *Genome Inform.* 2005 **16**: 95 [PMID: 16362911]
[12] Oh B *et al. Development.* 2000 **127**: 3795 [PMID: 10934024]
[13] Ogura K *et al. Development.* 2003 **130**: 2495 [PMID: 12702662]
[14] Gilleard JS & McGhee JD, *Mol Cell Biol.* 2001 **21**: 2533 [PMID: 11259601]
[15] Aburatani S, *Gene Regul Syst Bio.* 2011 **5**: 75 [PMID: 22272062]
[16] Sulston JE *et al. Dev Biol.* 1983 **100**: 64 [PMID: 6684600]
[17] Baugh LR *et al. Development.* 2005 **132**: 1843 [PMID: 15772128]
[18] Lei H *et al. Development.* 2009 **136**: 1241 [PMID: 19261701]
[19] Yanai *et al. Mol Sys Biol.* 2008 **4**: 163 [PMID: 18277379]
[20] Baugh LR *et al. Genome Biol.* 2005 **6**: R45 [PMID: 15892873]

Edited by P Kangueane
Citation: Aburatani, Bioinformation 8(14): 652-657 (2012)

BIOINFORMATION

open access

Supplementary material:

Table 1: Relationships among estimated factors and stage-specific expressio

		Factor loading			Detected stage		
Gene	**Communality**	**Factor 1 (F1)**	**Factor 2 (F2)**	**Factor 3 (F3)**	**Epidermal**	**Muscle**	**Early embryo**
tbx-8	.858	.857	-.350	.022			○(initiator)
tbx-9	.831	-.656	.631	-.043			○(initiator)
elt-1	.495	.682	.150	-.082	○		○
unc-120	.647	.754	-.278	-.041		○	
scrt-1	.311	.556	-.036	.009	○	○	○
vab-7	.775	-.590	.175	.267	○	○	
nob-1	.450	.674	-.563	-.055	○	○	
pop-1	.597	.493	-.586	.107		○	○
hnd-1	.750	.002	.866	-.017		○	
hlh-1	.466	-.343	-.024	.590		○	
lin-26	.732	-.018	.289	.805	○		
nhr-25	.415	-.125	.133	-.618	○		
elt-3	.413	-.060	-.096	.633	○		

Communality indicates the percent of variance in each gene, explained by the factors. Factor loading is the correlation coefficients between genes and factors. The red-colored number indicates the highest absolute value for each gene. The "detected stage" indicates the gene detected in the cell type and the developmental phase in embryogenesis. The detected stages were described in Yanai *et al.* **[19].**