Inference of Transcriptional Network for Pluripotency in Mouse Embryonic Stem Cells

S Aburatani
Team Leader, Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology, 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. In embryonic stem cells, various transcription factors (TFs) maintain pluripotency. To gain insights into the regulatory system controlling pluripotency, I inferred the regulatory relationships between the TFs expressed in ES cells. In this study, I applied a method based on structural equation modeling (SEM), combined with factor analysis, to 649 expression profiles of 19 TF genes measured in mouse Embryonic Stem Cells (ESCs). The factor analysis identified 19 TF genes that were regulated by several unmeasured factors. Since the known cell reprogramming TF genes (Pou5f1, Sox2 and Nanog) are regulated by different factors, each estimated factor is considered to be an input for signal transduction to control pluripotency in mouse ESCs. In the inferred network model, TF proteins were also arranged as unmeasured factors that control other TFs. The interpretation of the inferred network model revealed the regulatory mechanism for controlling pluripotency in ES cells.

1. Introduction
Clarification of the mechanism of pluripotency in mouse embryonic stem cells (ESCs) is an important theme in biology and regenerative medicine. Four transcription factors with pluripotency ability have been found, and these factors are known as the “Yamanaka factors” [1,2]. Since gene regulatory networks determine cell differentiation and specification [3], in order to decipher the mechanism of the pluripotency process in ESCs, the structures of the regulatory networks between genes should be elucidated. Although the regulatory networks for the genes encoding the important factors for pluripotency have been investigated [4,5,6], the global mechanism of pluripotency in ESCs has remained unclear. Many genes function in ESCs to control cell differentiation, and their expression is controlled by transcription factors (TFs). Therefore, an analysis of the regulatory network between TFs, which are related to pluripotency in mouse ESCs, will reveal the framework of the regulatory mechanism during the pluripotency process.

In order to clarify the regulatory network between TFs, I applied Structural Equation Modeling to huge expression profiles, which were measured in mouse ESCs. In previous investigations, the regulatory relationships between some TFs and their target genes were identified by ChIP-seq technology, gene expression data, and other experimental approaches. According to the former
investigations, some aspects of the regulatory mechanism between TFs were clarified [7,8]. In a previous investigation, I utilized an improved SEM approach to infer the regulatory networks during developmental stages in model organisms [9,10]. In my SEM approach, both the previous results and un-measured factors can be included in the network model.

In this study, I improved the method for assuming the initial model for SEM calculations, for inferring the regulatory network between TFs in mouse ESCs. I applied my approach to 19 TF genes with regulated expression in mouse ESCs. Using my method, four signal-transduction pathways were estimated as regulators of initiators for pluripotency, and some TFs were identified as being regulated by these signal transduction pathways.

2. Materials & Methods

2.1. Expression data and selected genes

We compiled the expression profiles measured in mouse ESCs from the GEO Database (http://www.ncbi.nlm.nih.gov/geo/). Among the 1211 Series data (GSE) in GEO, four Series data (GSE16375, GSE26520, GSE42135 and GSE43597) were selected as the expression profiles measuring the expression of genes under various conditions: 342 conditions in GSE16375, 246 conditions in GSE26520, 42 conditions in GSE42135 and 26 conditions in GSE43507. For the SEM calculation, the downloaded data were normalized by a logarithmic transformation and the Z score for each condition, and the normalized data for 656 conditions were combined as a matrix. From the empirically detected transcription factors in mouse/human ESCs, 19 transcription factors are known to be important in ESCs. In this study, I utilized these 19 transcription factor genes to infer the regulatory network between them.

2.2. Factor Analysis

In the SEM calculation, un-measured regulatory factors can be included in the network mode as latent variables. In order to assume the network model, the optimal number of latent variables in the model was estimated by performing factor analysis. By using factor analysis, the variability among the expression profiles of the 19 TF genes can be described by a linear combination of a smaller number of latent variables and error variables, as follows:

\[ x_i - u_i = \sum_{l=1}^{n} \alpha_l f_l + \epsilon_i \]

where \( x_i \) is the vector of the expression levels of gene \( i \), and \( u_i \) is the mean of \( x_i \). Note that \( n \) is the number of latent variables \( f \), \( \alpha_l \) is the regression weight of \( f_l \), and \( \epsilon_i \) is an independent error term with zero mean and finite variance. Linear models expressed by equation (1) were constructed for each gene, and arranged in a matrix form. In this matrix form, the variance-covariance matrix \( S \) between genes can be expressed by the structurized matrix \( M \), and each element of matrix \( M \) is expressed by parameters. A comparison between \( S \) and \( M \) allows the values of the parameters, defined by equation (1), to be estimated. In order to clarify the possible number of latent variables, the exploratory factor analysis with the Kaiser criterion states was performed. In the Kaiser criterion, the number of latent variables is equal to the number of eigenvalues of the covariance matrix that are greater than one. Furthermore, the causal relationships from the estimated latent variables to 19 TF genes were estimated by confirmatory factor analysis. The confirmatory factor analysis was performed with an assumed model with all possible causalities between estimated variables, and the relevant causalities were selected from the absolute values of the factor loadings.

2.3. Exploratory modeling for the initial model

According to the empirical knowledge of the regulatory relationships between TFs, three TFs are considered to be the initiators of sequential transcriptional regulation in ESCs. To construct an initial model that includes this knowledge within the network model, the three latent variables were arranged
as translated transcription factor proteins: POU5F1(Oct4), NANOG and SOX2 [1,2,3]. These latent variables are the targets that control their corresponding genes. In order to estimate the effects of these three latent variables on the other TF genes, I improved my developed four-step procedure for exploratory model assumption [9].

In our four-step procedure, all combinations of the regulatory relationships from one latent variable to the other TF genes were tested by SEM. As one of the constraints of the SEM calculation, at least one regulatory arrow from the latent variable to the other variables should be fixed. In this case, the fixed regulation was not tested. In this study, I constructed models with different fixed regulatory arrows, and applied the four-step procedure to determine the best fitting model structure under the restriction. The structures of the fitted models were compared to estimate the remaining regulatory relationships from the latent variable to the other TF genes. This process was applied to all three latent variables. Finally, the remaining regulatory relationships from the latent variables to the other TF genes were combined, to construct an initial model for the SEM calculation.

### 2.4. Structural Equation Modelling (SEM)

After the estimation of the relationships between the variables for an initial model, I applied the SEM calculation to determine the structure of the network model. According to the initial model, all variables were defined by linear models, as follows:

\[
I = AI + Bg + h \\
g = GI + Gg + e
\]

in which \(I\) is a vector of \(p\) latent variables; \(g\) is a vector of \(q\) measured gene expressions, \(A\) is a \(p \times p\) matrix representing the regulatory weight between latent variables, \(B\) is a \(p \times q\) matrix of effectiveness from the gene to the latent variables, \(G\) is a \(q \times p\) matrix representing the regulation from the latent variables to the genes, and \(I\) is a \(q \times q\) matrix representing the relationships between the genes. The error terms are denoted by the vectors \(h\) and \(e\), respectively. From these linear models, the covariance matrix of model \(S_m\), which is the matrix-valued function of the parameters, is given by

\[
S_m = G \left[ I - A \quad -B \right]^{-1} Cov, \left[ I - A \quad -B \right]^{-1} G'
\]

where \(G\) denotes a vector combined with \(p\) zero elements and \(q\) one elements, and \(Cov\) denotes the covariance matrix of the error terms. SEM is based on a covariance analysis to estimate the parameters for fitting \(S_m\) to the covariance matrix \(S\), which is calculated from the observed data. The SEM software package SPSS AMOS 21.0 (IBM, USA) and the R package (sem) were used for the SEM calculation. To optimize the network model, I applied my developed iteration algorithm to the initial model [9,10].

### 3. Results & Discussion

According to the factor analysis, 4 factors were estimated as regulatory factors for 19 TF genes in mouse ESCs. The 19 TF genes were divided into 4 clusters by their regulators. Furthermore, the well known pluripotency-related genes are regulated by different factors: Pou5f1 is regulated by 2 factors (F1 and F4), Sox2 is regulated by another factor (F2), and Nanog is regulated by the remaining factor (F3). In general, these genes are considered to be controlled by LIF and Wnt signal transduction. Thus, the estimated regulatory factors were regarded as the effects of signal transduction.

The four factors were arranged as latent variables in the initial model, and the regulatory relationships between factors and genes were also displayed as directed arrows, depending on their factor loading values. In order to include some knowledge within the initial model, I arranged three more latent variables in the initial model. These three latent variables indicated the TF proteins of POU5F1, NANOG and SOX2, which are known to be initiators of pluripotency by regulating other TF genes. By my developed exploratory modeling for the initial model, the effects of the three initiators were estimated: the POU5F1 protein regulated 10 other TF genes, the NANOG protein regulated 8
other TF genes, and the SOX2 protein regulated only 3 TF genes. These estimated effects from the three TF proteins were also included within the initial model.

The inferred network model is displayed in Figure 1a. Interestingly, the latent variable of the SOX2 protein was deleted from the network model during the model optimization process. Furthermore, the estimated regulatory factors (F1, F2, F3 and F4) were not related to each other. Figure 1b shows the schema of the inferred network. The strong associations between variables (weight >= 0.5) were extracted from the inferred network. Since F1 was a regulator of Pou5f1 expression, F1 was considered to be the effect of LIF signal transduction. In the lower class in F1, the Pou5f1 gene affected the Nanog gene and the POU5F1 protein. Actually, Pou5f1 is known to regulate Nanog gene expression. Furthermore, the effect from POU5F1 to Sox2 in the inferred model agreed with the known function of POU5F1.

Figure 1. Inferred networks. (a) Whole network model. (b) Network with strong associations. Blue: estimated factors, pink: known factors for pluripotency, yellow: important TFs in ESCs

4. Conclusion
Here, I inferred the transcriptional regulation between 19 TFs, which are considered to be important for pluripotency in mouse ESCs. By including unobserved variables, the signal transduction effects were clarified. Our network model revealed the hierarchical regulation by well known initiators for pluripotency.

5. References
[1] Takahashi K and Yamanaka S 2006 Cell 126 663
[2] Okita K, Ichisaka T, Yamanaka S 2007 Nature 448 313
[3] Davidson E H 2006 The Regulatory Genome: Gene Regulatory Networks in Development and Evolution (Burlington, MA: Academic Press)
[4] Jaenisch R and Young R 2008 Cell 132 567
[5] Zeng F, Baldwin D A and Schultz R M 2004 Dev. Biol. 272 483
[6] Xie D, Chen C C, Ptaszek L M, Xiao S, Cao X, Fang F, Ng H H, Lewin H A, Cowan C and Zhong S 2010 Genome Research 20 804
[7] Boyer L A, Lee T I, Cole M F, Johnstone S E, Levine S S, Zucker J P, Guenther M G, Kumar R M, Murray H L, Jenner R G, Gifford D K, Melton D A, Jaenisch R and Young R A 2005 Cell 122(6) 947
[8] Kim J, Chu J, Shen X, Wang J and Orkin S H 2008 Cell 132 1049
[9] Aburatani S 2012 Bioinformation 8(14) 652
[10] Aburatani S, Fujibuchi W, Yamane J, Nagano R, Sone H, Imanish S and Ohsako S 2012 Int. J. on Adv. in Life Sciences 5 103