REFERENCES

Ahn, S. M., Byun, K., Kim, D., Lee, K., Yoo, J. S., Kim, S. U., Jho, E. H., Simpson, R. J., and Lee, B. (2008). Olig2-induced neural stem cell differentiation involves downregulation of Wnt signaling and induction of Dickkopf-1 expression. *PLoS ONE* 3, e3917. doi:10.1371/journal.pone.0003917

Aiba, K., Sharov, A. A., Carter, M. G., Foroni, C., Vescovi, A. L., and Ko, M. S. (2006). Defining a developmental path to neural fate by global expression profiling of mouse embryonic stem cells and adult neural stem/progenitor cells. *Stem Cells* 24, 889–895.

Akamatsu, W., Deveale, B., Okano, H., Cooney, A. J., and van der Kooy, D. (2009). Suppression of Oct-4 by germ cell nuclear factor restricts pluripotency and promotes neural stem cell development in the early neural lineage. *J. Neurosci.* 29, 2113–2124.

Asano, H., Aonuma, M., Sanosaka, T., Kohyama, J., Namihira, M., and Nakashima, K. (2009). Astrocyte differentiation of neural precursor cells is enhanced by retinoic acid through a change in epigenetic modification. *Stem Cells* 27, 2744–2752.

Catena, R., Tiveron, C., Ronchi, A., Porta, S., Ferri, A., Tatangelo, L., Cavallaro, M., Favaro, R., Ottolenghi, S., Reinbold, R., Schöler, H., and Nicolis, S. K. (2004). Conserved POU binding DNA sites in the Sox2 upstream enhancer regulate gene expression in embryonic and neural stem cells. *J. Biol. Chem.* 279, 41846–41857.

Cosgaya, J. M., Garcia-Villalba, P., Perona, R., and Aranda, A. (1996). Comparison of the effects of retinoic acid and nerve growth factor on PC12 cell proliferation, differentiation, and gene expression. *J. Neurochem.* 66, 89–98.

Davis, A. P., Murphy, C. G., Rosenstein, M. C., Wiegers, T. C., and Mattingly, C. J. (2008). The Comparative Toxicogenomics Database facilitates identification and understanding of chemical-gene-disease associations: arsenic as a case study. *BMC Med. Genomics* 1, 48. doi:10.1186/1755-8794-1-48

Ellinger-Ziegelbauer, H., Fostel, J. M., Aruga, C., Bauer, D., Boitier, E., Deng, S. B., Dickinson, D., Le Fevre, A. C., Fornace, A. J., Grenet, O., Gu, Y. Z., Hoflack, J. C., Shiiyama, M., Smith, R., Snyder, R. D., Spire, C., Tanaka, G., and Aubrechtk, J. (2009). Characterization and interlaboratory comparison of a gene expression signature for differentiating genotoxic mechanisms. *Toxicol. Sci.* 110, 341–352.

Engberg, N., Kahn, M., Petersen, D. R., Hansson, M., and Serup, P. (2010). Retinoic acid synthesis promotes development of neural progenitors from mouse embryonic stem cells by suppressing endogenous, Wnt-dependent nodal signaling. *Stem Cells* 28, 1498–1509.

Graham, V., Khudyakov, J., Ellis, P., and Pevny, L. (2003). SOX2 functions to maintain neural progenitor identity. *Neuron* 39, 749–765.

Harnisha, D. C., Baruad, A. B., Sopranob, K. J., and Soprano, D. R. (1990). Induction of β-retinoic acid receptor mRNA by teratogenic doses of retinoids in murine fetuses. *Differentiation* 45, 103–108.

Hubal, E. A. C. (2009). Biologically relevant exposure science for 21st century toxicity testing. *Toxicol. Sci.* 111, 226–232.

Jin, Z., Liu, L., Bian, W., Chen, Y., Xu, G., Cheng, L., and Jing, N. (2009). Different transcription factors regulate nestin gene expression during P19 cell neural differentiation and central nervous system development. *J. Biol. Chem.* 284, 8160–8173.

Jukkola, T., Lahti, L., Naserke, T., Wurst, W., and Partanen, J. (2006). FGF regulated gene-expression and neuronal differentiation in the developing midbrain-hindbrain region. *Dev. Biol.* 297, 141–157.

Kawasaki, H., Mizuseki, K., Nishikawa, S., Kaneko, S., Kuwana, Y., Nakanishi, S., Nishikawa, S. I., and Sasai, Y. (2000). Induction of midbrain dopaminergic neurons from ES cells by stromal cell-derived inducing activity. *Neuron* 28, 31–40.

Kohonen, T. (1990). The self-organizing map. *Proc. IEEE* 78, 1464–1480.

Lamoury, F. M. J., Croitoru-Lamoury, J., and Brew, B. J. (2006). Undifferentiated mouse mesenchymal stem cells spontaneously express neural and stem cell markers Oct-4 and Rex-1. *Cytotherapy* 8, 228–242.

Lee, D. C., Hsu, Y. C., Chung, Y.-F., Hsiao, C. Y., Chen, S. L., Chen, M.-S., Lin, H. K., and Chiu, I.-M. (2009). Isolation of neural stem/progenitor cells by using EGF/FGF1 and FGF1B promoter-driven green fluorescence from embryonic and adult mouse brains. *Mol. Cell. Neurosci.* 41, 348–363.

Ligon, K. L., Kesari, S., Kitada, M., Sun, T., Arnett, H. A., Alberta, J. A., Anderson, D. J., Stiles, C. D., and Rowitch, D. H. (2006). Development of NG2 neural progenitor cells requires Olig gene function. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7853–7858.

Loh, Y. H., Wu, Q., Chew, J. L., Vega, V. B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., Wong, K. Y., Sung, K. W., Lee, C. W., Zhao, X. D., Chiu, K. P., Lipovich, L., Kuznetsov, V. A., Robson, P., Stanton, L. W., Wei, C. L., Ruan, Y., Lim, B., and Ng, H. H. (2006). The Oct-4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* 38, 431–440.

Maden, M., Sonneveld, E., van der Saag, P. T., and Gale, E. (1998). The distribution of endogenous retinoic acid in the chick embryo: implications for developmental mechanisms. *Development* 125, 4133–4144.

Mic, F. A., Molotkov, A., Benbrook, D. M., and Duester, G. (2003). Retinoid activation of retinoic acid receptor but not retinoid X receptor is sufficient to rescue lethal defect in retinoic acid synthesis. *Proc. Natl. Acad. Sci. U.S.A.* 100, 7135–7140.

Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003). The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES Cells. *Cell* 113, 631–642.

Miyazaki, K., Narita, N., and Narita, M. (2005). Maternal administration of thalidomide or valproic acid causes abnormal serotonergic neurons in the offspring: implication for pathogenesis of autism. *Int. J. Dev. Neurosci.* 23, 287–297.

Nagano, R., Akanuma, H., Qin, X. Y., Imanishi, S., Toyoshiba, H., Yoshinaga, J., Ohsako, S., and Sone, H. (2012). Multi-parametric profiling network based on gene expression and phenotype data: a novel approach to developmental neurotoxicity testing. *Int. J. Mol. Sci.* 13, 187–207.

Nishimura, F., Yoshikawa, M., Kanda, S., Nonaka, M., Yokota, H., Shiroi, A., Nakase, H., Hirabayashi, H., Ouji, Y., Birumachi, J., Ishizaka, S., and Sakaki, T. (2003). Potential use of embryonic stem cells for the treatment of mouse parkinsonian models: improved behavior by transplantation of in vitro differentiated dopaminergic neurons from embryonic stem cells. *Stem Cells* 21, 171–180.

Okada, Y., Shimazaki, T., Sobue, G., and Okano, H. (2004). Retinoic-acid-concentration-dependent acquisition of neural cell identity during in vitro differentiation of mouse embryonic stem cells. *Dev. Biol.* 275, 124–142.

Okazawa, H., Okamoto, K., Ishino, F., Ishinokaneko, T., Takeda, S., Toyoda, Y., Muramatsu, M., and Hamada, H. (1991). The Oct3 gene, a gene for an embryonic transcription factor, is controlled by a retinoic acid repressible enhancer. *EMBO J.* 10, 2997–3005.

Qin, X. Y., Wei, F., Yoshinaga, J., Yonemoto, J., Tanokura, M., and Sone, H. (2011). siRNA-mediated knockdown of aryl hydrocarbon receptor nuclear translocator 2 affects hypoxia-inducible factor-1 regulatory signaling and metabolism in human breast cancer cells. *FEBS Lett.* 585, 3310–3315.

Schuldiner, M., Eiges, R., Eden, A., Yanuka, O., Itskovitz-Eldor, J., Goldstein, R. S., and Benvenisty, N. (2001). Induced neuronal differentiation of human embryonic stem cells. *Brain Res.* 913, 201–205.

Scotland, K. B., Chen, S., Sylvester, R., and Gudas, L. J. (2009). Analysis of Rex1 (zfp42) function in embryonic stem cell differentiation. *Dev. Dyn.* 238, 1863–1877.

Seiler, A. E., Buesen, R., Visan, A., and Spielmann, H. (2006). Use of murine embryonic stem cells in embryotoxicity assays: the embryonic stem cell test. *Methods Mol. Biol.* 329, 371–395.

Shi, W., Wang, H., Pan, G., Geng, Y., Guo, Y., and Pei, D. (2006). Regulation of the pluripotency marker Rex-1 by Nanog and Sox2. *J. Biol. Chem.* 281, 23319–23325.

Shiotsugu, J., Katsuyama, Y., Arima, K., Baxter, A., Koide, T., Song, J., Chandraratna, R. A., and Blumberg, B. (2004). Multiple points of interaction between retinoic acid and FGF signaling during embryonic axis formation. *Development* 131, 2653–2667.

So, P. L., Yip, P. K., Bunting, S., Wong, L. F., Mazarakis, N. D., Hall, S., Mcmahon, S., Maden, M., and Corcoran, J. P. (2006). Interactions between retinoic acid, nerve growth factor and sonic hedgehog signalling pathways in neurite outgrowth. *Dev. Biol.* 298, 167–175.

Tanaka, S., Kamachi, Y., Tanouchi, A., Hamada, H., Jing, N., and Kondoh, H. (2004). Interplay of SOX and POU factors in regulation of the Nestin gene in neural primordial cells. *Mol. Cell. Biol.* 24, 8834–8846.

Thomas, R. S., Allen, B. C., Nong, A., Yang, L., Bermudez, E., Clewell,

H. J., and Andersen, M. E. (2007). A method to integrate benchmark dose estimates with genomic data to assess the functional effects of chemical exposure. *Toxicol. Sci.* 98, 240–248.

Tomioka, M., Nishimoto, M., Miyagi, S., Katayanagi, T., Fukui, N., Niwa, H., Muramatsu, M., and Okuda, A. (2002). Identification of Sox-2 regulatory region which is under the control of Oct-3/4-Sox-2 complex. *Nucleic Acids Res.* 30, 3202–3213.

Toyoshiba, H., Sone, H., Yamanaka, T., Parham, F. M., Irwin, R. D., Boorman, G. A., and Portier, C. J. (2006). Gene interaction network analysis suggests differences between high and low doses of acetaminophen. *Toxicol. Appl. Pharmacol.* 215, 306–316.

Toyoshiba, H., Yamanaka, T., Sone, H., Parham, F. M., Walker, N. J., Martinez, J., and Portier, C. J. (2004). Gene interaction network suggests dioxin induces a significant linkage between aryl hydrocarbon receptor and retinoic acid receptor beta. *Environ. Health Perspect.* 112, 1217–1224.

van Dartel, D. A. M., Pennings, J. L. A., Van Schooten, F. J., and Piersma, A. H. (2010). Transcriptomics-based identification of developmental toxicants through their interference with cardiomyocyte differentiation of embryonic stem cells. *Toxicol. Appl. Pharmacol.* 243, 420–428.

Veiga Quemelo, P. R., Lourenco, C. M., and Peres, L. C. (2007). Teratogenic effect of retinoic acid in swiss mice. *Acta Cir. Bras.* 22, 451–456.

Wilson, L., and Maden, M. (2005). The mechanisms of dorsoventral patterning in the vertebrate neural tube. *Dev. Biol.* 282, 1–13.

Yamanaka, T., Toyoshiba, H., Sone, H., Parham, F. M., and Portier, C. J. (2004). The TAO-Gen algorithm for identifying gene interaction networks with application to SOS repair in *E. coli*. *Environ. Health Perspect.* 112, 1614–1621.

Yang, H., Xia, Y., Lu, S. Q., Soong, T. W., and Feng, Z. W. (2008). Basic fibroblast growth factor-induced neuronal differentiation of mouse bone marrow stromal cells requires FGFR-1, MAPK/ERK, and transcription factor AP-1. *J. Biol. Chem.* 283, 5287–5295.

Zhang, Y. J., Xuan, J. H., de los Reyes, B. G., Clarke, R., and Ressom, H. W. (2008). Network motif-based identification of transcription factor-target gene relationships by integrating multi-source biological data. *BMC Bioinformatics* 9, 203. doi:10.1186/1471-2105-9-203

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 27 March 2012; accepted: 12 July 2012; published online: 07 August 2012.

Citation: Akanuma H, Qin X-Y, Nagano R, Win-Shwe T-T, Imanishi S, Zaha H, Yoshinaga J, Fukuda T, Ohsako S and Sone H (2012) Identification of stage-specific gene expression signatures in response to retinoic acid during the neural differentiation of mouse embryonic stem cells. Front. Gene. 3:141. doi: 10.3389/fgene.2012.00141

This article was submitted to Frontiers in Toxicogenomics, a specialty of Frontiers in Genetics.

APPENDIX

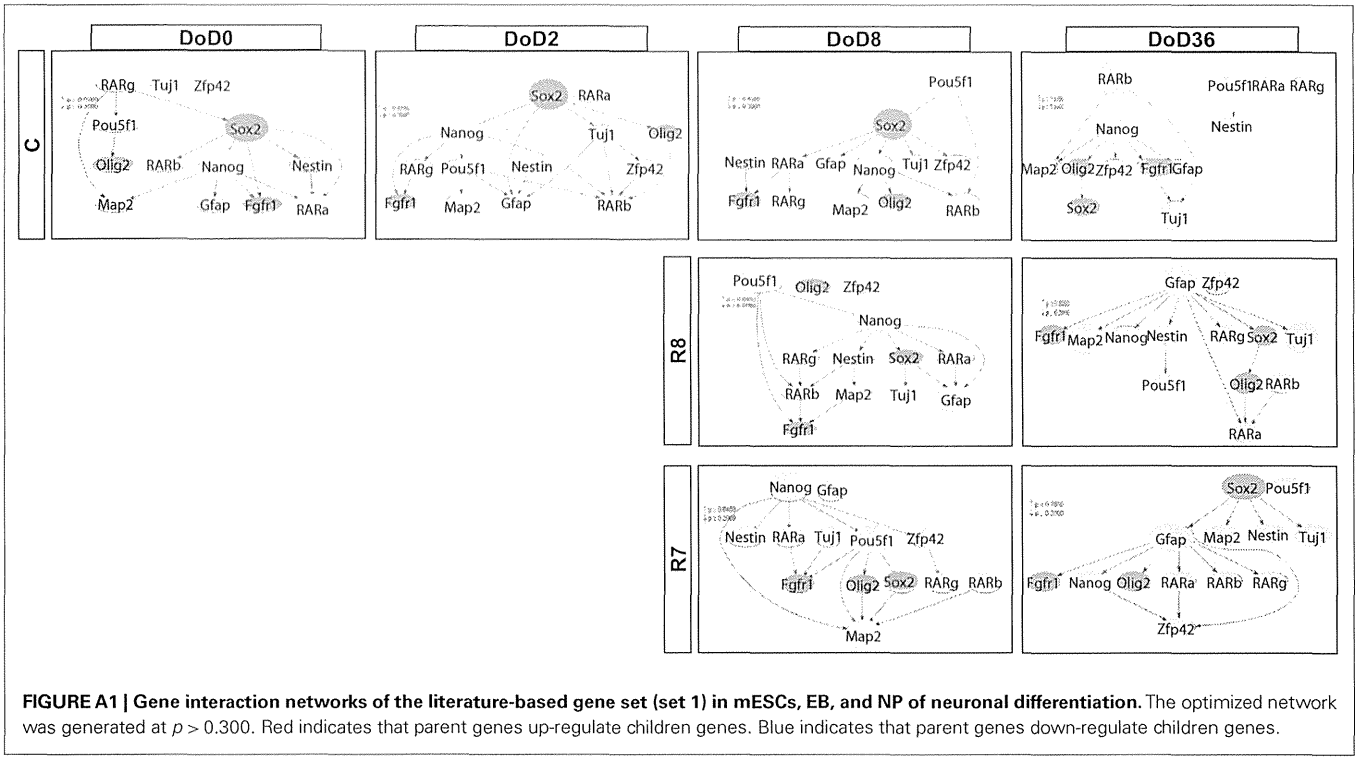


FIGURE A1 | Gene interaction networks of the literature-based gene set (set 1) in mESCs, EB, and NP of neuronal differentiation. The optimized network was generated at $p > 0.300$. Red indicates that parent genes up-regulate children genes. Blue indicates that parent genes down-regulate children genes.

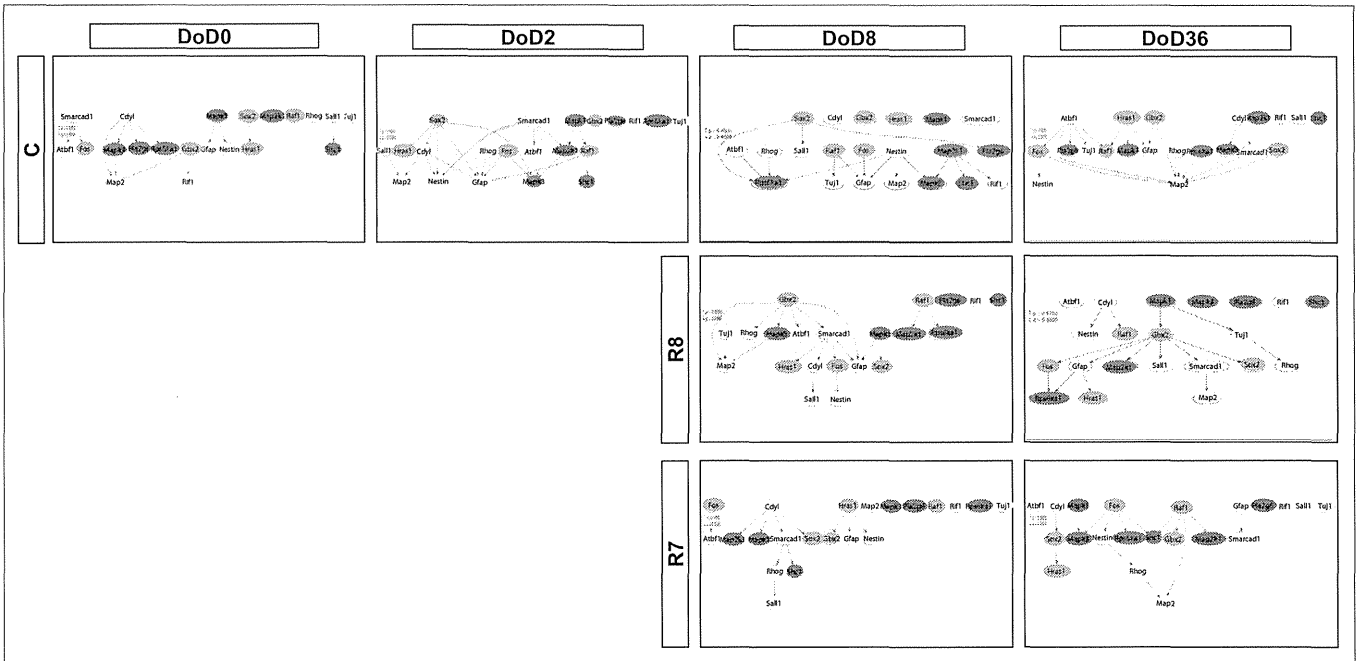


FIGURE A2 | Gene interaction networks of the analysis-based gene set (set 2) in mESCs, EB, and NP of neuronal differentiation. The optimized network was generated at $p > 0.300$. Red indicates that parent genes up-regulate children genes. Blue indicates that parent genes down-regulate children genes.

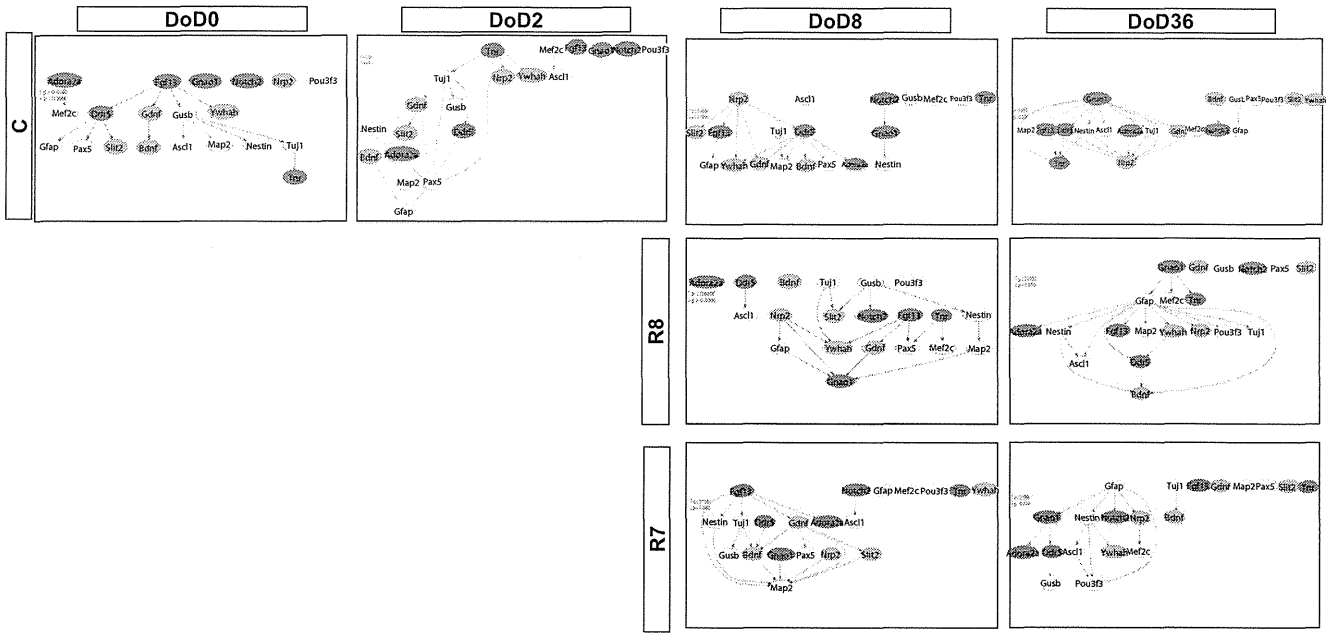


FIGURE A3 | Gene interaction networks of the enrichment gene set (set 3) in mESCs, EB, and NP of neuronal differentiation. The optimized network was generated at $p > 0.300$. Red indicates that parent genes up-regulate children genes. Blue indicates that parent genes down-regulate children genes.

Journal of Molecular Biology Research; Vol. 2, No. 1; 2012
ISSN 1925-430X E-ISSN 1925-4318
Published by Canadian Center of Science and Education

Inference of Specific Gene Regulation by Environmental Chemicals in Human Embryonic Stem Cells

Sachiyo Aburatani[1] & Wataru Fujibuchi[2]

[1] Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology, Tokyo, Japan

[2] Center for iPS Research and Application, Kyoto University, Kyoto, Japan

Correspondence: Sachiyo Aburatani, 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. Tel: 81-335-998-712. E-mail: s.aburatani@aist.go.jp

Received: December 2, 2012 Accepted: December 17, 2012 Online Published: December 27, 2012

doi:10.5539/jmbr.v2n1p54 URL: http://dx.doi.org/10.5539/jmbr.v2n1p54

Abstract

We are exposed to many environmental chemicals in our daily life. Certain chemicals threaten our health, especially that of embryos and can cause serious developmental problems. To prevent abnormal development and diseases caused by chemicals, it is important to clarify the mechanisms of chemical toxicity in embryonic cells. The gene regulatory network is one of the useful methods for clarifying functional mechanisms in living cells, so we applied a statistical method to infer the gene regulatory network in human embryonic stem cells. In this study, we improved our previously developed SEM approach for inferring a network model from 9 gene expression profiles in human embryonic stem cells, which were exposed to various chemicals. The estimated regulatory models clarified the differences between chemicals, and the shapes of the inferred models reflected the features of the chemical toxicities. The toxicity of acrylamide affected neuronal cell-related genes, while that of diethylnitrosamine disturbed cell differentiation-related genes. On the other hand, the TCDD network reflected feedback regulation, and finally disturbed neuronal cell-related genes. In the Thalidomide network, cell differentiation genes related to axis formation in embyronic cells were affected by thalidomide toxicity.

Keywords: structural equation modeling, environmental chemical, gene regulatory network, embryonic stem cell

1. Introduction

1.1 Introduction of the Problem

Environmental pollution is a byproduct of our usual life activities. Vehicle exhaust contains gases, including many noxious chemicals. Factories discharge industrial waste in the air, ground, and water. Many rivers are polluted by domestic sewage and wastewater. The emitted chemicals are sometimes trapped in clouds and then contaminate the ground in rainfall. Thus, we are exposed to many chemicals in our daily life, and some environmental chemicals can cause serious developmental toxicity effects. Developmental toxicity is either a structural or functional alteration, and these alterations interfere with the normal developmental programming in early embryos. These interferences can cause abnormal development and diseases (Baccarelli & Bollati, 2009; Hou et al., 2012). One of the most infamous environmental chemicals is methylmercury, which is known to affect fetal development (Yuan, 2012; Tatsuta et al., 2012). Furthermore, other chemicals are also considered to be toxic, since they can cause abnormal cell differentiation in embryos (Rappolee et al., 2012; He et al., 2012; Harrill et al., 2011).

To prevent chemically-induced developmental abnormalities and diseases, it is important to clarify the mechanisms of chemical toxicity in embryonic cells (Gündel et al., 2007; Thompson & Bannigan, 2008). The gene regulatory network is one of the useful methods to clarify the regulatory mechanisms. To infer the networks among the genes from the mRNA levels, various algorithms, including Boolean and Bayesian networks, have been developed (Akutsu et al., 2000; Friedman et al., 2000). In our previous investigation, we developed an approach based on graphical Gaussian modeling (GGM) in combination with hierarchical clustering, and we could infer the huge network among all of the genes by this approach. (Aburatani et al., 2003; Aburatani &

Horimoto, 2005). However, GGM infers only the undirected graph, whereas the Boolean and Bayesian models infer the directed graph, which shows causality. Although all of these approaches are suitable for establishing the relationships among the genes, they cannot reveal the relationships between un-observed factors and genes, because of insufficient information in the gene expression profiles. To clarify the mechanisms of biological processes in living cells, un-observed factors, which affect the target gene's expression, should also be considered. Thus, an alternative approach that includes un-observed factors should be applied.

Recently, we developed a new statistical approach based on Structural Equation Modeling (SEM), to infer the protein-DNA interactions for gene transcriptional control from only the gene expression profiles, in the absence of protein information (Aburatani, 2011; 2012). We applied this approach to reveal the causalities within the well-studied transcriptional regulation system in yeast (Aburatani, 2011). The significant features of SEM are the inclusion of latent variables within the constructed model and the ability to infer the network, including the cycle structure. Furthermore, the SEM approach allows us to strictly evaluate the inferred model, by using fitting scores. The linear relationships between variables are assumed to minimize the differences between the model's covariance matrix and the calculated sample covariance matrix. Some fitting indices are defined for evaluating the model adaptability, and thus the most suitable model can be selected by SEM (Bollen, 1989; Duncan, 1975; Pearl, 2001).

Here, we applied the SEM approach to infer the regulatory relationships among 9 neurodevelopmentally-related genes. The expression profiles of these genes were measured in human embryonic stem cells exposed to four environmental chemicals. The chemicals are known to have harmful toxicities that affect the developmental process in human embryos. Thus, inferring the regulatory network among the developmentally-related genes will help us to reveal the mechanisms of toxicity-dependent responses in the embryo. Furthermore, we improved our SEM approach for assuming preliminary initial models from the time-series data. By using this new approach, we can construct an initial model for the SEM calculation in the absence of known regulatory interactions. The resulting gene expression data clarified the chemical-specific interactions among the developmentally-related genes.

2. Methods

2.1 Expression Data

We utilized the expression data that were measured to clarify the effects of environmental chemical exposure on neuronal differentiation (He et al., 2012; Fujibuchi et al., 2011). In these expression data, nine genes considered to be affected by chemicals were measured in human embryonic stem cells: GATA2, Lmx1A, MAP2, Nanog, Nestin, Nodal, Oct3/4, Pax6 and Tuj1 (He et al., 2012; Fujibuchi et al., 2011). The expression of beta-actin was also measured, as an internal control. The expression levels of these 10 genes were measured in human embryonic stem cells exposed to four chemicals: acrylamide, diethylnitrosamine, TCDD and thalidomide (He et al., 2012; Fujibuchi et al., 2011). The toxicities of these chemicals are different: acrylamide is neurotoxic, diethylnitrosamine is genotoxic, TCDD is carcinogenic, and thalidomide has other toxicity. The human embryonic cells were exposed to each chemical for several time periods: 24 hours, 48 hours, 72 hours and 96 hours. Each chemical was also tested at 5 concentrations: very low, low, medium, high and very high. The expression of the selected genes was measured twice under each condition by RT-PCR, and thus 160 (4 time periods x 5 concentrations x 2 repeats x 4 chemicals) expression patterns per gene were measured (Fujibuchi et al., 2011).

First, the expression level of each gene was normalized to the internal beta-actin control and averaged, as follows:

$$E_g = \frac{1}{N}\sum_{i=1}^{N} \log_2\left(\frac{e_g^i}{e_{bActin}^i}\right) \quad (1)$$

Here, N is the number of repeated experiments, e_g^i is the measured expression level of gene g under one set of conditions, and e_{bActin}^i is the beta-actin expression level measured under the same conditions. By dividing by the expression level of beta-actin, the intracellular expression level of each gene was normalized. To minimize the experimental error, the logarithms of the normalized expression data were obtained and averaged.

2.2 Extraction of Causalities from Expression Data

Usually, we assume an initial model from previous knowledge for the SEM calculation, but there are no defined regulations among the selected genes in this study. Thus, we had to construct an initial model of each chemical from the regulatory relationships between the gene pairs. To detect the regulatory relationships from the measured time series expression data, cross correlation coefficients were applied to the expression profiles. These cross correlation coefficients were calculated for each chemical and each concentration. Cross correlation is utilized as a measure of similarity between two waves in signal processing by a time-lag application, and it is also applicable to pattern recognition (Li & Caldwell, 1999). In a time series analysis, the cross correlation between two time series describes the normalized cross covariance function. Therefore, the range of cross correlation values is from -1 to +1. If we let $X_t = \{x_1, \cdots, x_N\}, Y_t = \{y_1, \cdots, y_N\}$ represent two time series datasets including N time points, then the cross correlation is given by

$$r_{xy} = \frac{\sum_{t=1}^{N} \{x_i - \bar{x}\}\{y_{i+d} - \bar{y}\}}{\sqrt{\sum_{t=1}^{N} \{x_i - \bar{x}\}^2} \sqrt{\sum_{t=1}^{N} \{y_{i+d} - \bar{y}\}^2}} \tag{2}$$

where d is the time-lag between variables X and Y. In this case, the expression profiles were measured at four time points, and thus three cross correlations of each gene pair were calculated with d=-1, 0, and 1.

2.3 Construction of the Initial Models

To infer the chemical-dependent regulatory networks, the differences between times and concentrations should be merged. In this study, we developed a new method for constructing an initial model of each chemical, with the merging of time and concentration conditions. Figure 1 shows the newly developed method. First, we constructed lag matrices to simplify the information from the time series data. The elements of the lag matrices were the time lags, which were defined for the calculation of the cross correlation. In this study, cross correlations were calculated with three lags, -1, 0, and +1. The absolute values of these three cross correlations were compared, and the lag value d with the highest absolute value was arranged as a matrix element. Lag matrices were constructed for each concentration, and thus five lag matrices were obtained for each chemical (Figure 1a).

In the next step, we merged the difference in the concentrations of each chemical. Binomial relationships were extracted from each lag matrix. For each chemical, there are five lag matrices according to the chemical concentration, and we considered that the chemical-specific relationships among the genes will be conserved in several lag matrices. If the same relationships existed in several lag matrices, then the binomial relationships were duplicated (Figure 1b).

We subsequently constructed one frequency matrix of binary relationships for each chemical. We counted the frequency of the appearance of relationships in binomial relationships. The number representing the frequency of each gene pair was arranged in this matrix, and thus the range was from 0 to 5 (Figure 1c). In the frequency matrix, we can merge the differences in the concentrations, since the elements of the frequency matrix indicate the information for the different concentrations. We selected the possible relationships from the frequency matrix. It is considered that a possible relationship would be indicated by its frequency of appearance. Thus, we selected the relationships with two or more values in the frequency matrix (Figure 1d). At the final step, an initial model was constructed with the selected possible relationships. By this approach, an initial model can include cyclic structures (Figure 1e).

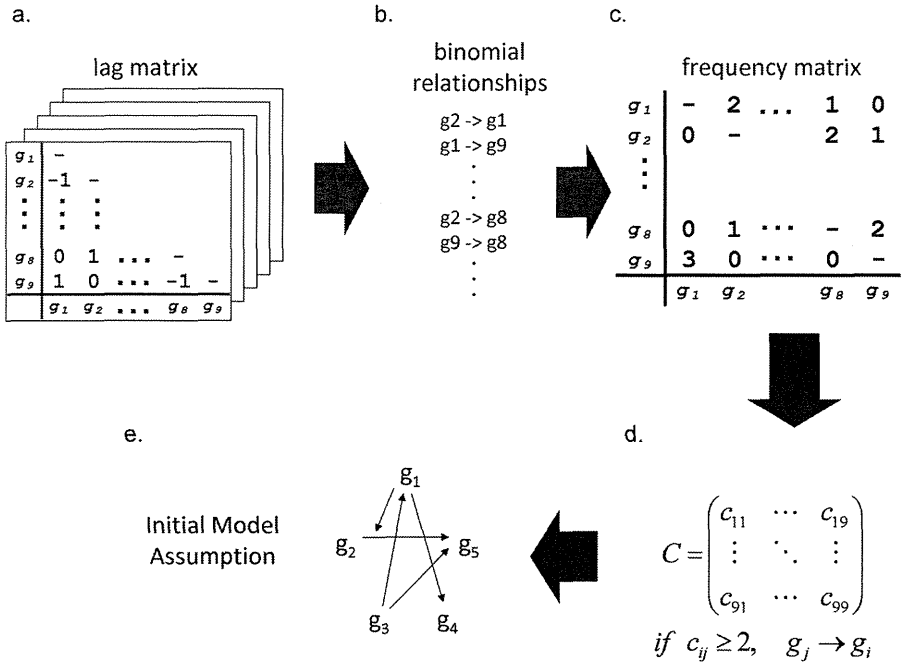


Figure 1. Developed procedure for initial model construction

The procedure for constructing an initial model from the time-lag information of the cross correlation coefficients. (a) Time-lag matrices for each chemical. In this study, three time-lags were selected for the calculation of the cross correlation coefficients. Thus, three cross correlation coefficient values were obtained between all gene pairs. The time-lag value with the highest absolute value among the cross correlation coefficients was selected. Time-lag matrices were constructed for each concentration, so five time-lag matrices were obtained for each chemical. (b) Binomial relationships. These relationships were extracted from the five time-lag matrices. If the same relationships exist in several concentration matrices, then the extracted binomial relationships are duplicated in this step. (c) Frequency matrix of causal relationships between all gene pairs. From the binomial relationship, we can count the frequency of relationships between gene pairs. (d) Selection of possible causal relationships from the frequency matrix. The possible relationships between genes are considered to persist at several chemical concentrations. Thus, we selected the relationships with two or more values in the frequency matrix. (e) Construction of an initial model with selected causal relationships. By this approach, an initial model can include cyclic structures.

2.4 Structural Equation Modeling without Latent Variables (SEM without LV)

After the construction of an initial model for each chemical, we applied the SEM calculation to infer the network model that fit the measured expression data. Usually, two types of variables can be included in the SEM model: observed and latent. These variables constitute the structural models that consider the relationships between the latent variables and the measurement models that consider the relationships between the observed variables and the latent variables. These relationships can be presented both algebraically, as a system of equations, and graphically, as path diagrams.

In this study, the nine developmentally-related genes (GATA2, Lmx1A, MAP2, Nanog, Nestin, Nodal, Oct3/4, Pax6 and Tuj1) were defined as the observed variables. Meanwhile, none were defined as latent variables, which were common regulators of several genes. The un-observed factor, which affected each gene's expression, was displayed as an error. The observed variables were classified as one of two types: exogenous variables and endogenous variables. Exogenous variables are not regulated by other variables in the system, as opposed to endogenous variables, which are regulated by other variables in the system. In the initial model, the starting genes are defined as exogenous variables without errors, while all other genes are defined as endogenous variables with errors. We inferred the regulatory relationships that exist between the observed variables in the network model. The model is defined as follows:

$$y = \Lambda y + \varepsilon \tag{3}$$

Here, y is a vector of p observed variables (measured gene expression patterns), and Λ is a $p \times p$ matrix representing the regulatory relationships between the observed variables. Errors that affect the observed endogenous variables are denoted by ε. The above equation can be represented in the SEM matrix format as:

$$\begin{bmatrix} O \\ y \end{bmatrix} = \begin{bmatrix} O & O \\ O & \Pi \end{bmatrix} \begin{bmatrix} O \\ y \end{bmatrix} + \begin{bmatrix} O \\ \varepsilon \end{bmatrix} \tag{4}$$

In this study, we did not define the latent variables, and thus Os were arranged as zero partial matrices, which denote no relationships with q latent variables. The SEM is based on a covariance analysis defined as $S = \Sigma(\theta)$, where S is the covariance matrix calculated from the observed data and $\Sigma(\theta)$ is the matrix-valued function of the parameter θ. Let Φ denote the covariance matrix of the error terms ε, and G denote the $p \times (p+q)$ combined matrix of the $p \times q$ zero matrix and the $p \times p$ identity matrix. The covariance matrix of model is given by

$$\Sigma(\theta) = G \begin{bmatrix} I - O & O \\ O & I - \Pi \end{bmatrix}^{-1} \Phi \begin{bmatrix} I - O & O \\ O & I - \Pi \end{bmatrix}^{-1'} G' \tag{5}$$

Each element of the covariance matrix model is expressed as a function of the parameters that appear in the model. The unknown parameters were estimated, in order to minimize the difference between the model covariance matrix and the sample covariance.

The SEM software package SPSS AMOS 17.0 (IBM, USA) was used to fit the model to the data. The quality of the fit was estimated by the goodness-of-fit index (GFI), which measures the relative discrepancy between the empirical data and the inferred model, and the adjusted GFI (AGFI), which is the GFI modified according to the degrees of freedom. Furthermore, we used CFI and RMSEA as fitting scores, to evaluate the model fitting. Since these indices have threshold values, as criteria to decide whether the model is suitable to obtain data independent of a huge sample number, they are considered to be useful to clarify the degree of model fitting in this study.

2.5 Parameter Estimation

Parameter estimation was performed by comparing the actual covariance matrix S, calculated from the measured data, and the estimated covariance matrices $\Sigma(\theta)$ of the constructed model. To minimize the difference between S and $\Sigma(\theta)$, the Maximum Likelihood (ML) method is commonly used as a fitting function to estimate the SEM parameters:

$$F_{ML}(S, \Sigma(\theta)) = \log|\Sigma(\theta)| - \log|S| + tr(\Sigma(\theta)^{-1} S) - p \tag{6}$$

Here, $\Sigma(\theta)$ is the estimated covariance matrix, S is the sample covariance matrix, $|\Sigma|$ is the determinant of matrix Σ, $tr(\Sigma)$ is the trace of matrix Σ, and p is the number of observed variables. The principal objective of SEM is to minimize $F_{ML}(S, \Sigma(\theta))$, which is the objective function and is used to obtain the maximum likelihood. Generally, $F_{ML}(S, \Sigma(\theta))$ is a nonlinear function. Therefore, iterative optimization is required to minimize $F_{ML}(S, \Sigma(\theta))$ and to find the solutions (Joreskog & Sorbom, 1984).

2.6 Iteration for Optimal Model

In the SEM analysis, both the parameters and network structures are fitted to the measured data. The parameters are estimated by maximum likelihood, and the network structures are evaluated by the scores of goodness of fit indices. The goodness of fit scores indicate the similarity between the constructed model and the measured data. Through the acceptance or rejection of the models, the optimal model that describes the measured data can be selected.

By using the estimated parameters, the variance-covariance matrix between the variables could be calculated in the network model. This model's variance-covariance matrix is compared with the actual variance-covariance matrix between observed variables, which is calculated from the measured data. The similarity between a constructed model and the actual data is defined in a quantitative manner by the fitting scores. In this study, four different fitting scores were utilized: GFI, AGFI, CFI and RMSEA. Values of GFI, AGFI and CFI above 0.90 are required for a good model fit. RMSEA is one of the most popular parsimony indexes displayed in the table, and RMSEA values below 0.05 represent a good model fit (Spirtes et al., 2001). Furthermore, RMSEA values of 0.10 or more are considered to indicate that the constructed model is far from the actual data. To optimize the model, we developed an iteration algorithm as follows:

Step 1: Deletion of a non-significant edge from the constructed network model

Use 0.05 as the significance level for the determination of the significant regulation among the variables. After the parameters are estimated, the inverse matrix of the Fisher information matrix of parameters is calculated. The inverse matrix of Fisher information represents the asymptotic parameters' covariance matrix. The probability of each parameter is calculated by using this asymptotic parameters' matrix, since all of the parameters are usually normally distributed.

Step 2: Reconstruction of the network model

The structure of the network model without the non-significant edge is completely different from that of the former model. Thus, all parameters should be re-calculated from the reconstructed model, and the similarity of the network structure should also be re-calculated.

Step 3: Iteration of Steps 1 and 2 until all edges become significant

Since the probabilities of all of the edges in the reconstructed models have also changed, the deletion of the non-significant edges is executed step-by-step.

Step 4: Addition of a possible causal edge to the reconstructed model

According to the Modification Index (MI), we add a new causal edge between the observed variables. The MI measures how much the chi-square statistic is expected to decrease if a particular parameter setting is constrained (Joreskog & Sorbom, 1984). The MI value indicates the possibility of new causality between the variables, and thus we add a new edge according to the highest MI score.

Step 5: Iteration from Steps 1 to 3

The addition of a new edge to a constructed model changes the structure of the network model again. In other words, all parameters, including the probabilities of all edges, have also changed. Thus, we execute the iteration from Step 1 to Step 3 again.

Step 6: Determination of significant relationships among error terms

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. The relationships between the error terms were considered to be other regulatory systems in the living cell. Thus, these relationships among the error terms were used for the calculations, but were not incorporated into the network, and thus they have been excluded from the figures.

3. Results

3.1 Initial Model Assumption

To construct the initial network model of each chemical, we utilized our newly developed method. One of the distinguishing features of our new method is its ability to include the cyclic structure in the network model. Cyclic regulation, such as feedback regulation, is considered to be important for living cells to control normal gene expression, and the new method is useful to detect the cyclic regulation from the gene expression data. The initially constructed models are shown in Figure 2. The initial model of TCDD was the most complex structure. The components of the constructed models were 9 genes with 19 relationships in Acrylamide, 8 genes with 12 relationships in Diethylnitrosamine, 9 genes with 23 relationships in TCDD, and 8 genes with 10 relationships in Thalidomide.

There are some obvious features in the network diagram of each initial model. The numbers of exogenous and endogenous genes are different from each other. In the initial Acrylamide model, four genes were arranged as exogenous variables, but only Oct3/4 was arranged as the last endogenous variable. Thus, it is considered that acrylamide quickly affected the expression of many genes, and only one gene was affected later. In contrast, only one gene was arranged as an exogenous variable and many genes were arranged as the last endogenous variables in the initial Thalidomide model. These differences between the initial chemical models summarized the distinctive gene expression profiles for each chemical. The initial TCDD model involved some cyclic regulation, even though the other models had only hierarchical regulation.

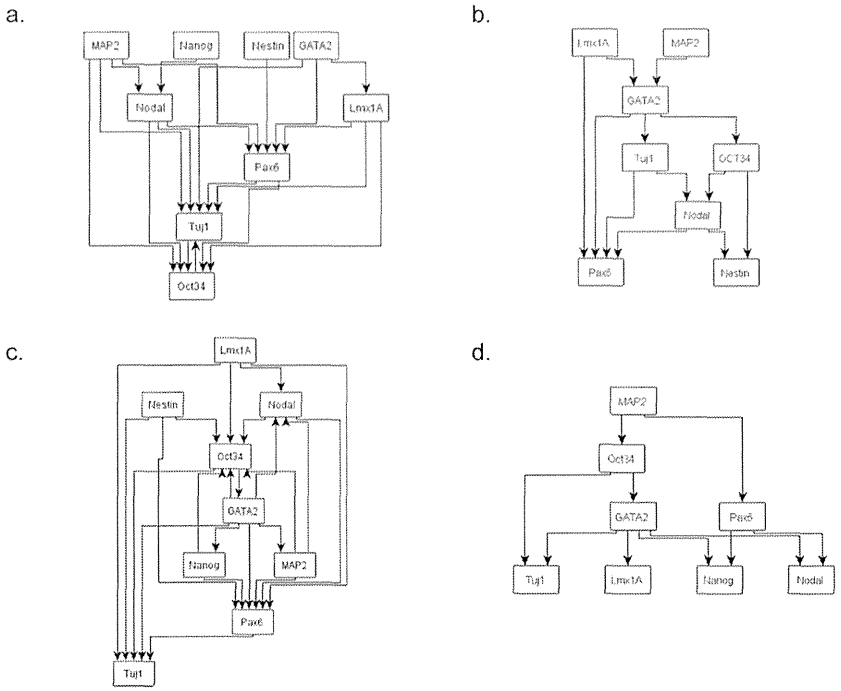


Figure 2. Initial network models

The initial models of the selected chemicals were constructed by the developed approach. One initial model was constructed for each chemical, since the initial model included summarized time-series information and concentration information. (a) Initial model constructed from all gene expression profiles with all Acrylamide exposure. (b) Initial model of Diethylnitrosamine. (c) Initial model of TCDD. (d) Initial model of Thalidomide. The numbers of genes in the initial models were 9 in Acrylamide, 8 in Diethylnitrosamine, 9 TCDD, and 8 in Thalidomide.

Before the calculation of SEM, all of the initial models were simplified, since the initial models included some duplicated interactions among the genes, such as direct interactions between two genes and indirect interactions between them. In the simplification process for the initial models, the longest path between two genes was retained, since the arrows indicated only time precedence, not causalities in the initial model. Therefore, the difference between direct and indirect interactions is not important. By retaining the longest paths, all of the preceding information was included, as the simplest diagram.

3.2 Inferred Networks by SEM

The final inferred networks for each chemical and the goodness of fit scores are depicted in Figure 3, and the estimated regression weights of the edges are displayed in Table 1. The inferred networks of the chemicals revealed distinct structures. The differences between the gene regulation by chemicals were clarified by the shapes of the inferred network models. The Acrylamide network was a centralized model, the Diethylnitrosamine network was a ladder-like model, the TCDD network was a closed circular structure, and the Thalidomide network was a diffusion type.

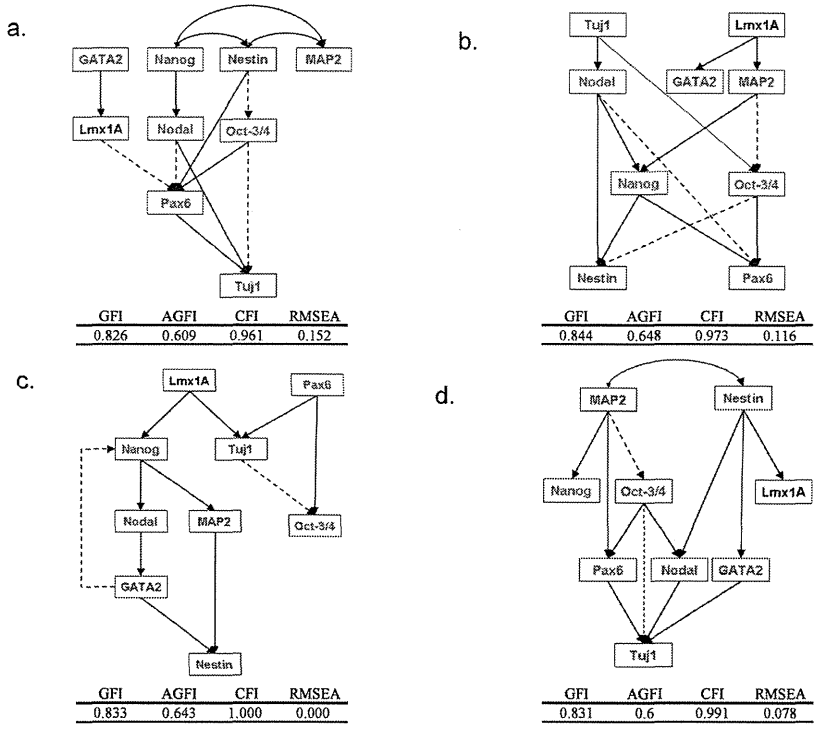


Figure 3. Inferred Toxic-dependent Networks

The optimal model for each chemical, obtained by the developed SEM iteration procedure. A positive relationship between genes is displayed with a solid arrow. A negative relationship between genes is displayed with a dashed arrow. Gene names with blue characters indicate "neurodevelopment related genes", genes with red characters indicate "cell differentiation-related genes" and genes with black characters indicate "related to transcription of insulin". (a) Acrylamide model; (b) Diethylnitrosamine model; (c) TCDD model and (d) Thalidomide model. The fitting scores are displayed under each model.

One of the unique features of the inferred Acrylamide network was that many genes were arranged at the top phase in the regulatory network, and only one gene was arranged as the final result of all regulation in the network. On the other hand, the shape of the Diethylnitrosamine network looked like a ladder, and two serial regulations interacted with each other. One serial regulation started from Lmx1A, and the other started from Tuj1. These top phase genes were considered as signal input genes, and they were different from those in the Acrylamide and Thalidomide networks. For example, Tuj1 was arranged as a signal input gene in the Diethylnitrosamine network, but it was arranged as an output object in the Acrylamide and Thalidomide networks. The unique feature of the TCDD network is the involvement of some closed circular structures in the inferred model. Among the parts of the circular structure, the regulatory direction from GATA2 to Nodal was different from the other relationships. Furthermore, the regression weight between GATA2 and Nodal was estimated as a negative value. Thus, it was considered that the inferred regulation from GATA2 to Nodal reflected feedback control by GATA2. In the Thalidomide network, the shape of the network model was reversed, as compared to that of the Acrylamide network. Only two genes were arranged at the top phase in the regulatory network, but many genes were arranged at the middle phase in the model. This means that only a few genes are directly affected by thalidomide, but finally many genes are affected throughout the gene regulatory network.

Table 1. Regression weight and probability of each edge

Acrylamide				Diethylnitrosamine			
Parent	Child	Regression Weight	P	Parent	Child	Regression Weight	P
GATA2	Lmx1A	0.921	***	Tuj1	Nodal	0.702	***
Nanog	Nodal	0.522	0.003	Lmx1A	MAP2	0.378	0.003
Nestin	Oct-34	-0.437	0.01	Tuj1	Oct-34	0.63	***
Nestin	Pax6	0.64	***	MAP2	Oct-34	-0.475	***
Nodal	Pax6	-0.803	***	Nodal	Nanog	0.295	***
Lmx1A	Pax6	-0.232	***	MAP2	Nanog	0.754	***
Oct-34	Pax6	0.592	***	Lmx1A	GATA2	0.636	***
Nodal	Tuj1	0.843	***	Nodal	Nestin	0.33	***
Pax6	Tuj1	1.09	***	Nodal	Pax6	-0.209	0.005
Oct-34	Tuj1	-0.702	***	Nanog	Pax6	0.418	***
				Nanog	Nestin	0.902	***
				Oct-34	Pax6	1.11	***
				Oct-34	Nestin	-0.193	***

TCDD				Thalidomide			
Parent	Child	Regression Weight	P	Parent	Child	Regression Weight	P
GATA2	Nanog	-0.787	***	MAP2	Oct-34	-0.443	0.023
GATA2	Nestin	0.22	***	MAP2	Pax6	0.349	0.005
Lmx1A	Nanog	1.374	***	Nestin	Nodal	1.03	***
Lmx1A	Tuj1	0.476	0.004	Nestin	GATA2	0.664	***
MAP2	Nestin	0.906	***	Oct-34	Pax6	0.932	***
Nanog	MAP2	1.024	***	Oct-34	Nodal	0.258	***
Nanog	Nodal	0.967	***	Oct-34	Tuj1	-0.597	***
Nodal	GATA2	0.931	***	Pax6	Tuj1	1.12	***
Pax6	Oct-34	0.988	***	Nodal	Tuj1	0.349	***
Pax6	Tuj1	0.5	0.003	GATA2	Tuj1	0.167	0.015
Tuj1	Oct-34	-0.324	***	MAP2	Nanog	0.84	***
				Nestin	Lmx1A	0.842	***
				Tuj1	Nanog	0.196	0.002

4. Discussion

Our inferred model revealed the differences between the gene regulation by environmental chemicals. Furthermore, the shapes of the network models reflected the different features of the chemical toxicities well. In the Acrylamide network, the effects of acrylamide toxicity finally aggregated to Tuj1, which is known to contribute to microtubule stability in neuronal cells (Rosenstein et al., 2003). Acrylamide is neurotoxic, and thus it is reasonable that its effect finally aggregated to a neuronal cell-related gene.

In the Diethylnitrosamine network, the cell differentiation genes were arranged from the middle to lower steps. This means that diethylnitrosamine disturbed normal cell differentiation in the embryonic stem cell. These harmful effects were considered to be caused by the carcinogenic genotoxicity of diethylnitrosamine (Ito et al., 1992; Puatanachokchai et al., 2006; Iatropoulos et al., 2006). On the other hand, the neuronal-related genes were arranged at a later phase in the TCDD network model. Although both diethylnitrosamine and TCDD have the same carcinogenic toxicities, their regulatory mechanisms were different.

From the Thalidomide network, it was considered that the receptors of thalidomide toxicity may be rarer than those of other chemicals. However, several types of genes are finally affected by thalidomide chemical toxicity. Among the cell differentiation genes, Nodal and Nanog are important for normal early embryonic development. Nodal is related to the development of the left-right axial structure (Hamada et al., 2002; Grandel & Patel, 2009), and its signaling pathway is important very early in development, for cell fate determination and many other developmental processes (Grandel & Patel, 2009). Nanog is a key factor for maintaining pluripotency in embryonic stem cells (Mitsui, 2003; Chambers et al., 2003). According to the abnormal expression of these cell differentiation-related genes, the thalidomide phenotype, with its harmful side effects, may occur in the human embryo.

We applied an improved SEM approach to reconstruct a gene regulatory model from the gene expression data in human embryonic stem cells, and we have shown that SEM is a powerful approach to estimate the gene

regulation caused by chemical toxicity. The inferred networks clarified the differences between the gene regulation by chemicals, and the features of the chemical toxicities were well reflected in the network structures. Thus, the network construction by SEM is one of the useful approaches for inferring the regulatory relationships among genes. Furthermore, the inferred network among genes can be utilized for the estimation of a chemical's effect, from experimentally obtained expression profiles. The ability to identify expression profiles and the corresponding biological functions is expected to provide further possibilities for SEM in the inference of the effects of chemical toxicity on regulatory mechanisms.

Acknowledgements

We would like to express our gratitude to Dr. Yamane (Kyoto Univ.) and Dr. Ohsako (Univ. of Tokyo) for providing the expression profiles in human embryonic stem cells exposed to 15 chemicals. Their support was quite valuable for this investigation.

References

Aburatani, S. (2012). Network Inference of pal-1 Lineage-Specific Regulation in the C. elegans Embryo by Structural Equation Modeling. *Bioinformation, 8*(14), 652-657. http://dx.doi.org/10.6026/97320630008652

Aburatani, S., & Horimoto, K. (2005). Elucidation of the Relationships between LexA-Regulated Genes in the SOS response. *Genome Informatics, 16*(1), 95-105.

Aburatani, S., Kuhara, S., Toh, H., & Horimoto, K., (2003). Deduction of a gene regulatory relationship framework from gene expression data by the application of graphical Gaussian modeling. *Signal Processing, 83*, 777-788. http://dx.doi.org/10.1016/S0165-1684(02)00476-0

Aburatani. S. (2011). Application of structure equation modeling for inferring a serial transcriptional regulation in yeast. *Gene. Regul. Syst. Bio., 5*, 75-88. http://dx.doi.org/10.4137/GRSB.S7569

Akutsu, T., Miyano, S., & Kuhara, S. (2000). Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *J. Comput. Biol., 7*, 331-343. http://dx.doi.org/ 10.1145/332306.332317

Baccarelli, A., & Bollati, V. (2009). Epigenetics and environmental chemicals. *Curr. Opin. Pediatr., 21*(2), 243-251. http://dx.doi.org/10.1097/MOP.0b013e32832925cc

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: Wiley-Interscience.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., & Smith, A. (2003) Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell, 113*(5), 643-655. http://dx.doi.org/10.1016/S0092-8674(03)00392-1

Duncan, O. D. (1975). *Introduction to Structural Equation Models*. New York: Academic Press.

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol., 7*, 601-620. http://dx.doi.org/10.1145/332306.332355

Fujibuchi, W. (2011). Prediction of Chemical Toxicity by Network-based SVM on ES-cell Validation System. *The Proc. of the 2011 Joint Conference of CBI-Society and JSBi.*

Grandel, C., & Patel, N. H. (2009). Nodal signaling is involved in left-right asymmetry in snails. *Nature, 457*(7232), 1007-1011. http://dx.doi.org/10.1038/nature07603

Gündel, U., Benndorf, D., Bergen, M., Altenburger, R., & Küster, E. (2007). Vitellogenin cleavage products as indicators for toxic stress in zebra fish embryos: A proteomic approach. *Proteomics, 7*(24), 4541-4554. http://dx.doi.org/10.1002/pmic.200700381

Hamada, H., Meno, C., Watanabe, D., & Saijoh, Y. (2002). Establishment of vertebrate left-right asymmetry. *Nat. Rev. Genet., 3*(2), 103-113. http://dx.doi.org/10.1038/nrg732

Harrill, J. A., Robinette, B. L., & Mundy, W. R. (2011). Use of high content image analysis to detect chemical-induced changes in synaptogenesis in vitro. *Toxicol. In. Vitro., 25*(1), 368-387. http://dx.doi.org/10.1016/j.tiv.2010.10.011

He, X., Imanishi, S., Sone, H., Nagano, R., Qin, X-Y., Yoshinaga, J., ... Ohsako, S. (2012). Effects of methylmercury exposure on neuronal differentiation of mouse and human embryonic stem cells. *Toxicol. Lett., 212*, 1-10. http://dx.doi.org/10.1016/j.toxlet.2012.04.011

Hou, L., Zhang, X., Wang, D., & Baccarelli, A. (2012). Environmental chemical exposures and human epigenetics. *Int. J. Epidemiol., 41*(1), 79-105. http://dx.doi.org/10.1093/ije/dyr154

Iatropoulos, M. J., Wang, C. X., Keutz, K. E., & Williams, G. M. (2006). Assessment of chronic toxicity and carcinogenicity in an accelerated cancer bioassay in rats of Nifurtimox, an antitrypanosomiasis drug. *Exp. Toxicol. Pathol., 57*(5-6), 397-404. http://dx.doi.org/10.1016/j.etp.2006.01.005

Ito, N., Hasegawa, R., Imaida, K., Masui, T., Takahashi, S., & Shirai, T. (1992). Pathological markers for non-genotoxic agent-associated carcinogenesis. *Toxicol. Lett., 64-65*, 613-620. http://dx.doi.org/10.1016/0378-4274(92)90239-G

Joreskog, K. G., & Sorbom, D. (1984). LISREL-VI: Analysis of Linear Structural Relationships By the Method of Maximum Likelihood. Redondo Beach: Doss-Haus Books.

Li, L., & Caldwell, G. E. (1999). Coefficient of cross correlation and the time domain correspondence. *J. of Electromyography and Kinesiology, 9*, 385-389. http://dx.doi.org/10.1016/S1050-6411(99)00012-7

Mitsui, K. (2003). The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell, 113*(5), 631-642. http://dx.doi.org/10.1016/S0092-8674(03)00393-3

Pearl, J. (2001). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge: Cambridge University Press.

Puatanachokchai, R., Kakuni, M., Wanibuchi, H., Kinoshita, A., Kang, J. S., Salim, E. I., ... Fukushima, S. (2006). Lack of promoting effects of phenobarbital at low dose on diethylnitrosamine-induced hepatocarcinogenesis in TGF-alpha transgenic mice. *Asian Pac. J. Cancer Prev., 7*(2), 274-278.

Rappolee, D. A., Xie, Y., Slater, J. A., Zhou, S., & Puscheck, E. E. (2012). Toxic stress prioritizes and imbalances stem cell differentiation: implications for new biomarkers and in vitro toxicology tests. *Syst. Biol. Reprod. Med., 58*(1), 33-40. http://dx.doi.org/10.3109/19396368.2011.647381

Rosenstein, J. M., Mani, N., Khaibullina, A., & Krum, J. M. (2003). Neurotrophic effects of vascular endothelial growth factor on organotypic cortical explants and primary cortical neurons. *J. Neurosci., 23*(35), 11036-11044.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, Prediction, and Search* (2nd ed.). Cambridge: The MIT Press.

Tatsuta, N., Nakai, K., Murata, K., Suzuki, K., Iwai-Shimada, M., Yaginuma-Sakurai, K., ... Satoh, H. (2012). Prenatal exposures to environmental chemicals and birth order as risk factors for child behavior problems. *Environ. Res., 114*, 47-52. http://dx.doi.org/10.1016/j.envres.2012.02.001

Thompson, J., & Bannigan, J. (2008). Cadmium: toxic effects on the reproductive system and the embryo. *Reprod. Toxicol., 25*(3), 304-315. http://dx.doi.org/10.1016/j.reprotox.2008.02.001

Yuan, Y. (2012). Methylmercury: a potential environmental risk factor contributing to epileptogenesis. *Neurotoxicology, 33*(1), 119-126. http://dx.doi.org/10.1016/j.neuro.2011.12.014

Application of Structural Equation Modeling for Inferring Toxicity-Dependent Regulation in Human Embryonic Stem Cells

Sachiyo Aburatani,
Computational Biology Research Center
National Institute of AIST
Tokyo, Japan
s.aburatani@aist.go.jp

Wataru Fujibuchi
Center for iPS Research and Application
Kyoto University,
Kyoto, Japan
w.fujibuchi@cira.kyoto-u.ac.jp

Abstract—**Chemical toxicity threat our daily health, especially for embryos. Reveling toxicity-dependant regulation in human embryo is one of the effective approaches to prevent some chemical effects. In previous study, we developed a network inference approach, based on Structural Equation Modeling (SEM). In this study, we improved the SEM approach and applied this enhanced approach to expression profiles in human embryonic stem cells exposed to various chemicals. The inferred gene regulatory models among neurodevelopment related genes clarify the differences between chemicals, and the network shapes reflected the features of chemical toxicities. The effects of Acrylamide toxicity finally aggregated to a neuronal cell-related gene, even though Diethylnitrosamine disturbed normal cell differentiation-related genes. Furthermore, gene regulatory network with Thalidomide was complicated, but embryonic development-related genes were estimated as the finally effected genes by Thalidomide toxicity.**

Keywords-Structural Equation Modeling; Embryonic Stem Cell; Gene Regulatory Network; Chemical Toxicity.

I. INTRODUCTION

We are exposed to many chemicals in our daily life, and chemical toxicity is known to exert harmful effects on human health. Actually, some diseases are caused by exposure to environmental pollution [1][2], including chemicals such as methylmercury [3][4], and so on. Furthermore, some chemical toxins are threatening, since they can cause abnormal cell differentiation in embryos [5][6][7]. Clarifying the details of the toxic stress response in embryonic cells is crucial for the prevention of harmful chemical effects [8][9].

To gain a better understanding of the role of the toxic stress response, a gene regulatory network is useful. With the gene expression information, the regulatory networks among the genes can be inferred. Various algorithms, including Boolean and Bayesian networks, have been developed to infer complex functional gene networks [10][11]. In our former investigation, we developed an approach based on graphical Gaussian modeling (GGM). The GGM approach is combined with hierarchical clustering for calculations with massive amounts of gene expression data, and we can infer the huge network among all of the genes by this approach [12][13]. However, GGM infers only the undirected graph, whereas the Boolean and Bayesian models infer the directed graph, which shows causality.

Recently, we developed a new statistical approach, based on Structural Equation Modeling (SEM) in combination with factor analysis and a four-step procedure [14]. This approach allowed us to reconstruct a model of transcriptional regulation that involves protein-DNA interactions, from only the gene expression data. Furthermore, SEM approach allows us to strictly evaluate the inferred model by using fitting scores. The SEM approach is available for the detection of causality among selected genes, as the linear relationships between genes are assumed to minimize the difference between the fitted model covariance matrix and the calculated sample covariance matrix [15][16][17].

Here, we applied the SEM approach to the limited expression data of neurodevelopment related genes in human embryonic stem cells exposed to various chemicals. The chemicals were considered to be toxic and to adversely affect the neurodevelopment related genes. Thus, inferring the gene regulatory network among neurodevelopment related genes will help to elucidate the toxic stress response in the human embryo. Since the regulatory interactions among the genes were unclear, a new approach for assuming an initial model should be developed for the application of SEM. In this study, we used an improved SEM approach that includes a new method for constructing a preliminary initial model, in the absence of known regulatory interactions. The resulting gene expression data clarified the chemical-specific interactions among the neurodevelopment related genes.

II. MATERIAL AND METHODS

A. Expression data

We were provided the expression data which were measured in previous investigation [6], and the details of data are follows. The nine genes considered to be affected by chemicals were measured in the human embryonic stem cells: GATA2, Lmx1A, MAP2, Nanog, Nestin, Nodal, Oct3/4, Pax6 and Tuj1 [6][18]. As an internal control, the expression of beta-actin was also measured. The expression data were obtained from human embryonic stem cells exposed to 15 chemicals [6][18]. The toxicity of each chemical was classified into one of three types: Neurotoxic, Carcinogenic and others. The human embryonic cells were exposed to each chemical for several time periods: 24 hours,

48 hours, 72 hours and 96 hours. Each chemical was also tested at 5 concentrations: very low, low, middle, high and very high. The expression of the selected genes was measured twice under each condition by RT-PCR, and thus 300 expression patterns per gene were measured [18].

The measured expression level of each gene was normalized as follows:

$$E_g = \frac{1}{N}\sum_{i=1}^{N}\log_2\left(\frac{e_g^i}{e_{bActin}^i}\right) \quad (1)$$

Here, N is the number of repeated experiments, e_g^i is the measured expression level of gene g under one set of conditions, and e_{bActin}^i is the beta-actin expression level measured under the same conditions. The expression level of each gene was divided by that of beta-actin, for intracellular normalization. To minimize the experimental error, the logarithms of the normalized expression data were obtained and averaged.

B. *Extraction of causalities from expression data*

For the iteration of model fitting in SEM, an initial model should be assumed from known information. To construct the initial model among the 9 neurodevelopment genes from the time series expressions, we applied cross correlation to the expression profiles measured for each chemical and each concentration.

Cross correlation is utilized as a measure of similarity between two waves in signal processing by a time-lag application, and it is also applicable to pattern recognition [19]. The cross correlation values range between -1 and +1. In a time series analysis, the cross correlation between two time series describes the normalized cross covariance function. Let $X_t = \{x_1, \cdots, x_N\}$, $Y_t = \{y_1, \cdots, y_N\}$ represent two time series data including N time points, and then the cross correlation is given by

$$r_{xy} = \frac{\sum_{t=1}^{N}\{x_t - \bar{x}\}\{y_{t+d} - \bar{y}\}}{\sqrt{\sum_{t=1}^{N}\{x_t - \bar{x}\}^2}\sqrt{\sum_{t=1}^{N}\{y_{t+d} - \bar{y}\}^2}} \quad (2)$$

where d is the time-lag between variables X and Y. In this case, the expression profiles were measured at 4 time points, and thus three cross correlations of each gene pair were calculated with $d = -1, 0, 1$.

C. *Construction of the initial model*

In this study, we focused on the chemical-specific regulatory network, and thus the differences between times and concentrations could be merged for the construction of the initial model. Figure 1 shows the new method developed for constructing an initial model of each chemical, with the merging of several conditions. The time difference was summarized by the cross correlations among genes. The time lag, which was defined for the calculation of the cross correlation, was used for the extraction of causality between all gene pairs. According to the time lags, three cross correlations were calculated between each gene pair, and we compared them with the absolute values of the cross correlations. The value d, with the highest cross correlation, was selected as the causal information between the gene pairs, and a matrix composed of the selected ds was constructed as the time lag matrix of each chemical at one concentration. Thus, five time lag matrices were constructed for each chemical (Fig. 1a).

To obtain the chemical-specific interactions among genes, we extracted the binomial relationships between gene pairs from the five constructed time lag matrices for each chemical (Fig. 1b). From the binomial relationships, we constructed a frequency matrix for each chemical, composed of the frequencies of all gene pairs (Fig. 1c). In this step, the difference in the concentration is merged as the frequency in the matrix. We extracted the gene pairs with frequency matrix values greater than or equal to two, as the chemical-specific regulation (Fig. 1d). From the extracted relationships between the genes, we reconstructed an initial model for each chemical (Fig. 1e). These initial models included the time series information as the directions of edges, and the different concentrations of each chemical were summarized as the existence of edges in the model.

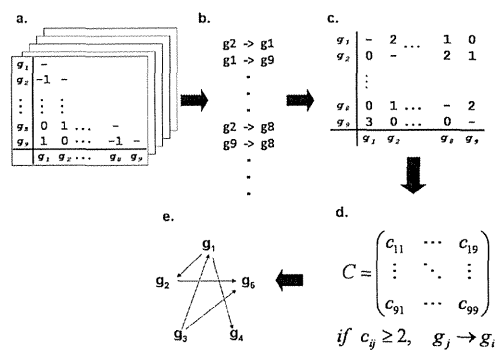


Figure 1. Developed procedure for initial model construction.

The procedure for constructing an initial model from the time-lag information of the cross correlation coefficients. (a) Time-lag matrices for each chemical. In this study, three time-lags were selected for the calculation of the cross correlation coefficients. Thus, three cross correlation coefficient values were obtained between all gene pairs. The time-lag value with the highest absolute value among the cross correlation coefficients was selected. Time-lag matrices were constructed for each concentration, so five time-lag matrices were obtained for each chemical. (b) Binomial relationships. These relationships were extracted from the five time-lag matrices. If the same relationships exist in several concentration matrices, then the extracted binomial relationships are duplicated in this step. (c) Frequency matrix of causal relationships between all gene pairs. From the binomial relationship, we can count the frequency of relationships between gene pairs. (d) Selection of possible causal relationships from the frequency matrix. The possible relationships between genes are considered to persist at several chemical concentrations. Thus, we selected the relationships with two or more values in the frequency matrix. (e) Construction of an initial model with selected causal relationships. By this approach, an initial model can include cyclic structures.

D. Structural Equation Modeling without Latent Variables (SEM without LV)

In general, SEM is a comprehensive statistical model that includes two types of variables: observed and latent. These variables constitute the structural models that consider the relationships between the latent variables and the measurement models that consider the relationships between the observed variables and the latent variables. These relationships can be presented both algebraically, as a system of equations, and graphically, as path diagrams.

In this study, the selected genes (GATA2, Lmx1A, MAP2, Nanog, Nestin, Nodal, Oct3/4, Pax6 and Tuj1), which are related to neurogenesis, were defined as the observed variables. Meanwhile, none were defined as latent variables. All observed variables were categorized into one of two types of variables, exogenous and endogenous, according to their interactions with other variables. Exogenous variables are those that are not regulated by the other variables, and endogenous variables are regulated by the others. In the initial model, the starting genes are defined as exogenous variables, while all other genes are defined as endogenous variables. Regulatory relationships exist between the observed variables in the network models. The model is defined as follows:

$$y = \Lambda y + \varepsilon \tag{3}$$

Here, y is a vector of p observed variables (measured gene expression patterns), and Λ is a $p \times p$ matrix representing the regulatory relationships between the observed variables. Errors that affect the observed endogenous variables are denoted by ε. The SEM software package SPSS AMOS 17.0 (IBM, USA) was used to fit the model to the data.

E. Parameter Estimation

Parameter estimation was performed by comparing the actual covariance matrix, calculated from the measured data, and the estimated covariance matrices of the constructed model. Maximum likelihood is commonly used as a fitting function to estimate SEM parameters:

$$F_{ML}(S, \Sigma(\theta)) = \log|\Sigma(\theta)| - \log|S| + tr(\Sigma(\theta)^{-1} S) - p \tag{4}$$

Here, $\Sigma(\theta)$ is the estimated covariance matrix, S is the sample covariance matrix, $|\Sigma|$ is the determinant of matrix Σ, $tr(\Sigma)$ is the trace of matrix Σ, and p is the number of observed variables. The principal objective of SEM is to minimize $F_{ML}(S, \Sigma(\theta))$, which is the objective function and is used to obtain the maximum likelihood. Generally, $F_{ML}(S, \Sigma(\theta))$ is a nonlinear function. Therefore, iterative optimization is required to minimize $F_{ML}(S, \Sigma(\theta))$ and to find the solutions [20].

F. Iteration for Optimal Model

The regulatory network analysis by SEM consists of two parts: parameter fitting and structure fitting. After the parameters of the constructed model are estimated by maximum likelihood, the network structures are evaluated according to the goodness of fit between the constructed model and the measured data. Through acceptance or rejection of the models, the optimal model that describes measured data can be selected.

In the network model, the covariance matrix between variables is calculated by the estimated parameters. The similarity between a constructed model and the actual relationships is predicted by comparing the matrix calculated from the network model to the matrix calculated from the actual data. To detect quantitative similarity between a constructed model and an actual relationship, fitting scores were developed. In this study, the quality of the fit was predicted by four different fitting scores: GFI, AGFI, CFI and RMSEA. Values of GFI, AGFI and CFI above 0.90 are required for a good model fit. RMSEA is one of the most popular parsimony indexes displayed in the table, and RMSEA values below 0.05 represent a good model fit [21]. Furthermore, RMSEA values of 0.10 or more are considered to indicate that the constructed model is far from the actual data.

To optimize the model, an iteration algorithm was developed, as follows:

Step1: Deletion of a non-significant edge from the model. Use 0.05 as the significance level for the determination of the chemical-specific interactions among genes. The output of SEM programs includes the probability of each edge, and thus we deleted the edge with the highest probability.

Step2: Reconstruction of the network model. The structure of the network model without the non-significant edge is different from the former network model. Thus, all parameters should be re-calculated from the reconstructed model, and the similarity of the network structure is also re-calculated.

Step3: Iteration of Steps 1 and 2 until all edges become significant. Since the probabilities of all of the edges in the reconstructed models have also changed, the deletion of the non-significant edges is executed step-by-step.

Step4: Addition of a possible causal edge to the reconstructed model. According to the Modification Index (MI), we add a new causal edge between the observed variables. The MI value indicates the possibility of new causality between the variables, and thus we add a new edge according to the highest MI score.

Step5: Iteration from Steps 1 to 3. By the addition of a new edge to a constructed model, the structure of network model is changed again. In other words, all parameters, including the probabilities of all edges, have also changed again. Thus, we execute the iteration from Step 1 to Step 3 again.

Step6: Determination of significant relationships among error terms. After all of the edges are significant and all of the MI scores are lower than 10.0 in the constructed model, significant relationships between 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.

III. Results and Discussion

A. The chemical concentrations have no effect

In this study, gene expression was measured in the presence of various chemicals, with several exposure times and at different concentrations. To reveal the most effective factor for gene expression, ANOVA and Tukey's HSD test were applied to the measured data. In statistics, ANOVA is utilized to detect differences between groups in terms of some variables. Since the chance of committing a type I error will be increased by performing multiple two-sample t-tests, a statistical test is needed to determine whether or not the means of more than two groups should be applied. The use of Tukey's HSD test clarified which means are significantly different from one another. Interestingly, the groups that were classified by the concentration of chemicals showed no significant difference in their gene expressions. Thus, the concentration of chemicals had no effect on the expression of the tested genes in the ES cells.

The numbers of significantly expressed genes between each chemical pair are shown in Table 1. From this table, the differences in the gene expression were not significant among the same type of toxic chemicals. Furthermore, the toxicity difference between neurotoxic and carcinogenic did not cause an expression difference for almost all of the genes. However, the exposure to 'other' chemicals, such as Thalidomide, bisphenol A and Permethrin, caused significant expression changes in many genes. To reveal the differences in gene expression due to the type of chemical toxicity, we selected one chemical, which was the most different from those of the other toxicity types, as the representative chemical for each toxicity type.

TABLE I. Number of Genes with Sinificantly Different Expression

		Methylmercury	2-Nitropropane	Acrylamide	p-Nitroaniline	4-hydoroxy PCB107	Benzo[a] pyrene	Diethylnitrosamine	Diethylaminofluorene	Phenobarbital	Tamoxifen	Diethylbestrol	TCDD	Thalidomide	Bisphenol-A	Permethrin
Neurotoxicity	Methylmercury	-														
	2-Nitropropane	1	-													
	Acrylamide	0	4	-												
	p-Nitroaniline	1	2	1	-											
	4-hydoroxy PCB107	2	4	0	0	-										
Carcinogenic	Benzo[a] pyrene	4	5	1	2	0	-									
	Diethylnitrosamine	4	3	5	4	5	5	-								
	Diethylaminofluorene	3	3	4	5	5	7	0	-							
	Phenobarbital	3	2	4	0	1	2	6	6	-						
	Tamoxifen	3	4	3	1	0	0	6	6	0	-					
	Diethylbestrol	3	5	2	0	1	0	5	6	0	0	-				
	TCDD	3	2	1	0	0	1	5	6	0	0	0	-			
others	Thalidomide	5	7	5	5	5	4	7	7	4	4	4	5	-		
	Bisphenol-A	5	6	5	5	3	4	6	7	6	3	3	3	0	-	
	Permethrin	4	6	4	5	4	4	7	7	3	4	4	4	0	0	-
		Methylmercury	2-Nitropropane	Acrylamide	p-Nitroaniline	4-hydoroxy PCB107	Benzo[a] pyrene	Diethylnitrosamine	Diethylaminofluorene	Phenobarbital	Tamoxifen	Diethylbestrol	TCDD	Thalidomide	Bisphenol-A	Permethrin
		Neurotoxicity					Carcinogenic							others		

Expression profiles were compared between all chemical pairs for each gene by Turkey's HSD test, and the genes that were estimated as exhibiting significantly different expression dependent on chemicals were counted.

The toxicities of the 15 chemicals were divided into 3 categories, Neural, Carcinogenic and others. Thus, three chemicals, Acrylamide, Diethylnitrosamine and Thalidomide, were selected as the representative toxic chemicals for neural, carcinogenic and others, respectively.

B. Genes are hierarchically controlled by chemical toxicity

We utilized the new method to construct the initial gene regulatory network models under the conditions with the three chemicals. Even though the new method can detect the cyclic interactions among genes, such as feed-back regulation, the structures of the constructed initial models indicated the hierarchical regulations among the genes. Figure 2 shows the constructed initial network models. The initial models of Acrylamide, Diethylnitrosamine and Thalidomide were composed of 9 genes with 19 relationships, 8 genes with 14 relationships, and 8 genes with 10 relationships, respectively.

There are some obvious features in the hierarchical diagram of each initial model. The numbers of exogenous and endogenous genes are different from each other. The initial Acrylamide model was composed of 4 genes as exogenous genes, but only Oct3/4 was the last endogenous gene. Thus, the expression profiles of Acrylamide indicated the quick responses of many genes after the chemical exposure, and only one gene was affected later. In contrast, Thalidomide exposure induced the expression of only one gene. These differences between the initial chemical models summarized the distinctive gene expression profiles for each chemical.

All of the initial models included some duplicated gene interactions, such as a direct interaction between two genes and an indirect interaction between them. To simplify these duplicated interactions, we only retained the longest path between two genes, since the regulation displayed by a direct path could be replaced by indirect paths in the model. In the initial model, the edges do not represent the direct regulation, but the preceding information. Thus, the difference between direct interaction and indirect interaction in the initial model is not very important. By retaining the longest paths, all of the preceding information was included, as the simplest diagram.

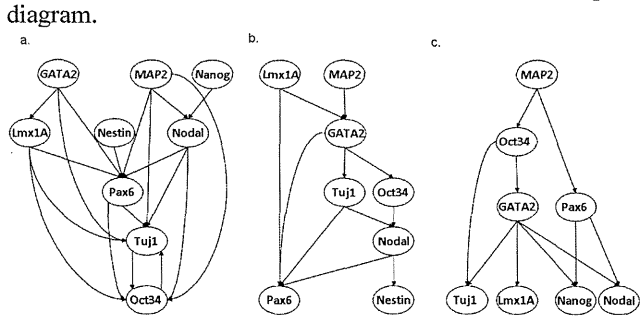


Figure 2. Initial network models.

The initial models which include summarized time-series information and concentration information. (a) Initial model of Acrylamide. (b) Initial model of Diethylnitrosamine. (c) Initial model of Thalidomide.

C. *Inferred Network by SEM*

The final inferred networks for each chemical and the estimated regression weights of the edges are depicted in Figure 3. The inferred networks of chemicals revealed distinct structures. In the inferred network of Acrylamide, many genes were arranged as exogenous objects, and only one gene was arranged as the final result of all regulation in the network. On the other hand, two serial regulations interacted with each other in the Diethylnitrosamine network model. One serial regulation was from Lmx1A to Pax6, and the other was from Tuj1 to Nestin. The signal input genes in the Diethylnitrosamine network were also different from those in the Acrylamide network. Even though Tuj1 was arranged as an output object in the Acrylamide network, Tuj1 was arranged as input in the Diethylnitrosamine network. The inferred network of Thalidomide was also different from both the Acrylamide and Diethylnitrosamine networks. In the Thalidomide network, only two genes were arranged as input objects, but four genes were arranged as output objects. This means that only a few genes will be directly affected by Thalidomide, but finally many genes were affected throughout the gene regulatory network.

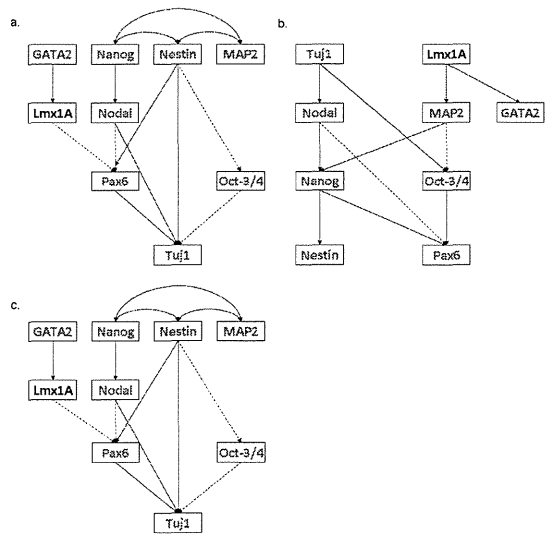
According to our inferred network, the differences between the gene regulation by chemicals were clarified, and the network shapes reflected the features of chemical toxicities. In the inferred network, the effects of Acrylamide toxicity finally aggregated to Tuj1, which is known to contribute to microtubule stability in neuronal cells [22]. Acrylamide is neurotoxic, and thus it is reasonable that the effect of Acrylamide finally aggregated to a neuronal cell-related gene.

As compared with the Acrylamide network, the cell differentiation genes were arranged at downstream steps in the Diethylnitrosamine network. From the carcinogenic features of Diethylnitrosamine [23][24][25], normal cell differentiation in the embryonic stem cell may be disturbed.

The most complicated structure was the Thalidomide network. In the Thalidomide network, several type of genes are finally affected by its chemical toxicity. Particularly, two cell differentiation-related genes, Nodal and Nanog, are important for normal early embryonic development. Nodal is related to the development of the left-right axial structure [26][27], and its signaling pathway is known to be important very early in development for cell fate determination and many other developmental processes [27]. Nanog is known as a key factor for maintaining pluripotency in embryonic stem cells [28][29]. Thus, the unusual expressions of these genes, which occurred due to Thalidomide toxicity, may have caused its harmful side effects.

IV. CONCLUSION

We applied an improved SEM approach to reconstruct a gene regulatory model from gene expression data in human embryonic stem cells, and we have shown that SEM is a powerful approach to estimate the gene regulation caused by chemical toxicity. The inferred networks clarified the differences between the gene regulation by chemicals, and the features of chemical toxicities were well reflected in the network structures. Thus, the network construction by SEM is one of the useful approaches for inferring the regulatory relationships among genes. Furthermore, the inferred



d.

Acrylamide			Diethylnitrosamine			Thalidomide		
parent	child	regression weight	parent	child	regression weight	parent	child	regression weight
GATA2	Lmx1A	0.919	Tuj1	Nodal	0.702	MAP2	Oct-34	-0.448
Nanog	Nodal	0.515	Lmx1A	MAP2	0.537	MAP2	Pax6	0.43
Nestin	Oct-34	-0.442	Tuj1	Oct-34	0.623	Oct-34	Pax6	0.951
Lmx1A	Pax6	-0.129	MAP2	Oct-34	-0.492	Nestin	Nodal	1.026
Nestin	Pax6	0.606	MAP2	Nanog	0.9	Nestin	GATA2	0.664
Nodal	Pax6	-0.844	Nodal	Nanog	0.284	Oct-34	Nodal	0.25
Oct-34	Pax6	0.643	Lmx1A	GATA2	0.636	Oct-34	Tuj1	-0.477
Pax6	Tuj1	1.145	Nodal	Pax6	-0.209	GATA2	Tuj1	0.183
Oct-34	Tuj1	-0.77	Nodal	Nestin	0.269	MAP2	Nanog	0.948
Nodal	Tuj1	0.933	Nanog	Nestin	0.843	Nestin	Lmx1A	0.831
MAP2	Nanog	0.968	Oct-34	Pax6	1.121	Pax6	Tuj1	1.002
MAP2	Nestin	0.944	Nanog	Pax6	0.428	Nodal	Tuj1	0.308
Nanog	Nestin	0.966				Nestin	MAP2	0.922

Figure 3. Inferred network by SEM.

The optimal model for each chemical, obtained by the developed SEM iteration procedure. A positive relationship between genes is displayed with a solid arrow. A negative relationship between genes is displayed with a dashed arrow. Gene names with blue characters indicate "neurodevelopment related genes", genes with red characters indicate "cell differentiation-related genes" and genes with black characters indicate "related to transcription of insulin". (a) Acrylamide model, (b) Diethylnitrosamine model and (c) Thalidomide model. (d) The estimated regression weights of all edges in the optimal models.