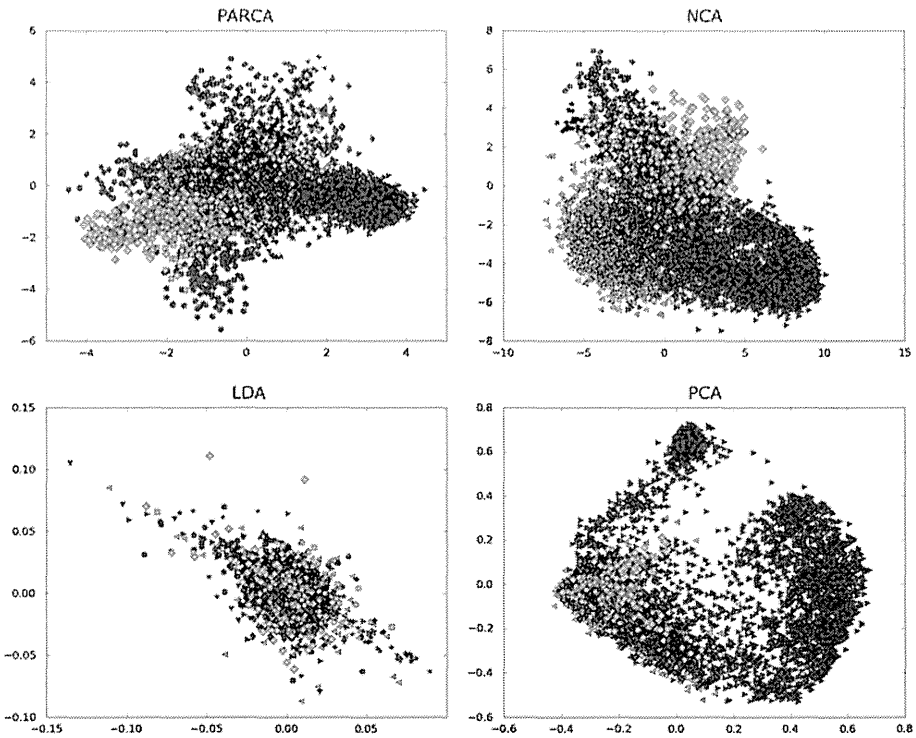


Fig. 7 Visualization results on the Reuters test set, for 2 % of training data. *Each colored symbol* represents one of the 13 classes (color figure online)

Table 2 Average running times (minutes)

Data set	LDA	ITML	NCA	PARCA	Log PARCA
Balance	0.01	0.20	134.70	1.79	4.63
Breast cancer	0.01	0.19	29.90	0.62	5.30
Ionosphere	0.01	0.07	3.59	0.26	5.49
Wine	0.01	0.01	4.26	0.03	0.77
20 Newsgroups	1.34	19.94	184.96	8.16	135.77
Reuters	0.98	9.39	33.91	4.46	128.70
WebKB	0.70	1.85	4.96	0.45	10.21
Binary	0.02	0.99	133.72	4.22	39.54
MNIST	0.63	3.18	81.95	7.05	145.08
UMIST	–	–	20.77	5.80	6.68
USPS	0.02	0.68	104.90	6.36	99.32

text data sets, 100 % of the data were used for training. For 20 Newsgroups, Reuters, and WebKB, 10 % of the data were used for training.

The result shows that LDA is the fastest method on all data sets except WebKB and UMIST data sets. Whereas NCA, PARCA, and Logistic PARCA are gradient-based methods, LDA involves an eigenvalue problem (see Sect. 2.1), which can be solved efficiently using

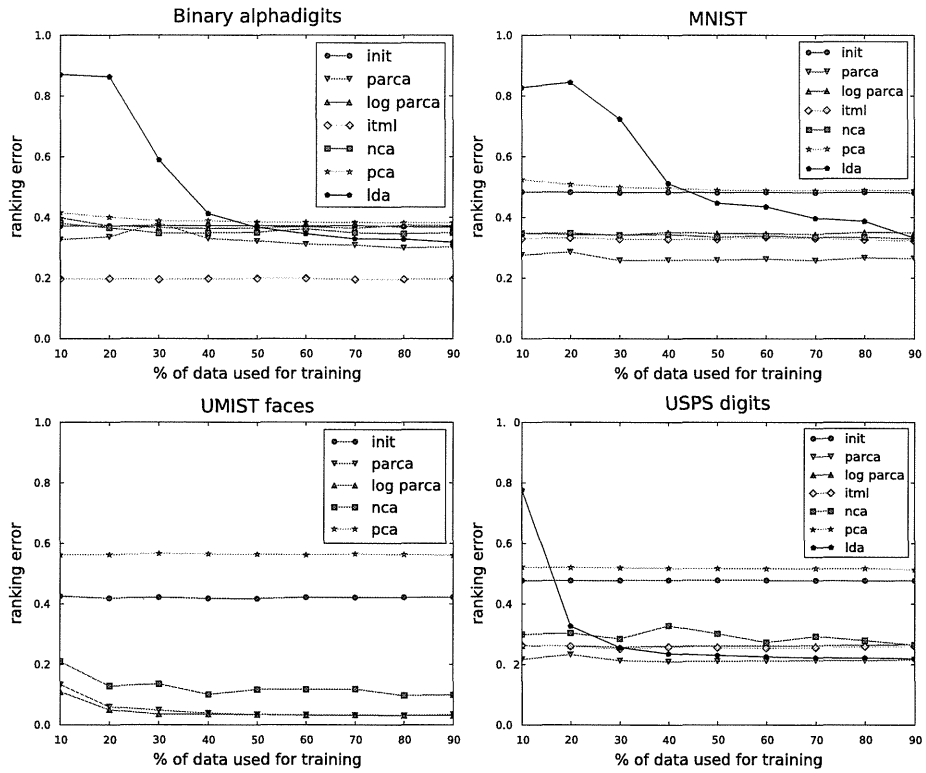


Fig. 8 Ranking performance on four image data sets: Binary alphadigits, MNIST handwritten digits, UMIST faces, and USPS handwritten digits

standard linear algebra libraries. Meanwhile, ITML is an iterative method that selects a random constraint at each iteration of the algorithm (see [11]). However, we empirically observe that ITML has a fast convergence rate. On three UCI data sets and three image data sets, ITML achieves the second lowest running time. In summary, LDA and ITML show the best running times on the four UCI data sets and three image data sets. However, their performance could not be assessed in the UMIST data set that has the highest dimensionality ($D = 10{,}304$). As we explained in Sect. 2.1 and 2.3, LDA and ITML involve the computation of matrices of size ($D \times D$) which is memory-demanding when D is large. Therefore, LDA and ITML are difficult to use when the dimensionality of the given data set is large.

We also observe that PARCA is the fastest method on the WebKB data set, and the second fastest on the 20 Newsgroups and Reuters data sets. Compared to Logistic PARCA, PARCA is significantly faster on all the data sets. As we have explained in Sect. 3.3.2, there are two basic differences between those two approaches: the use of the logistic function h to enforce non-negativity and the addition of the entropy function $T(h(\mathbf{B}))$ (Eq. 7). The running time differences can be accounted for by neither the additional computational complexity induced by the logistic function nor $T(h(\mathbf{B}))$, because it is a simple summation over $k = 1 \ldots K$ and $d = 1 \ldots D$. This suggests that the entropy function slows down the convergence of Logistic PARCA with respect to the number of iterations of the gradient descent. NCA and Logistic PARCA show the largest running times across all the data sets, which suggests that the convergence rate of the gradient descent method is slow.

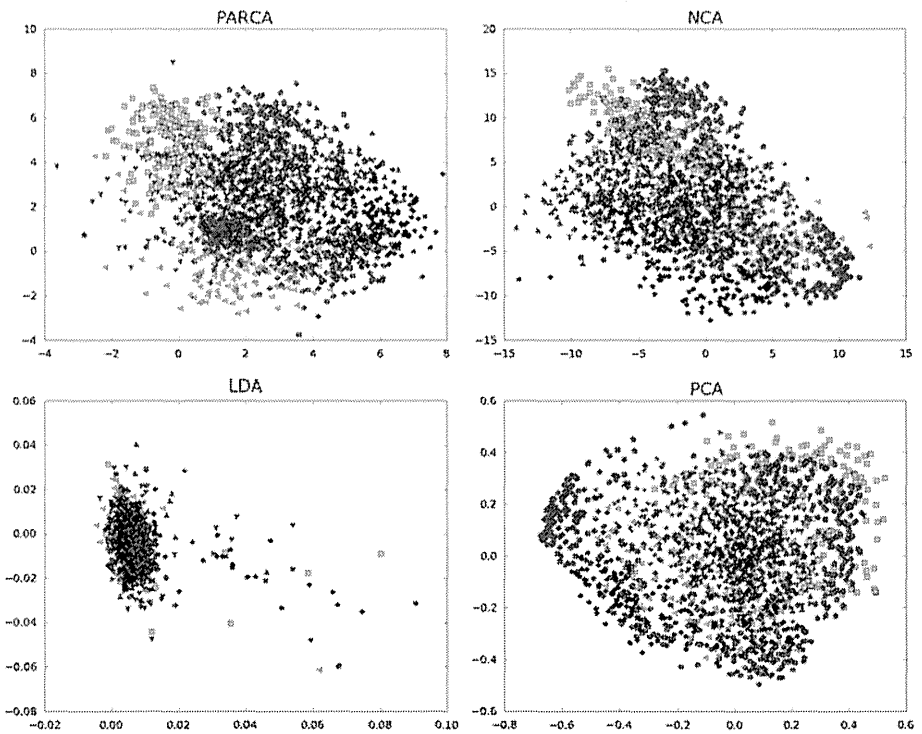


Fig. 9 Visualization results on the USPS handwritten digits, for 10 % of training data. *Each colored symbol* represents one of the ten classes (color figure online)

Now, we consider the memory complexity of PARCA and Logistic PARCA. As they are dimensionality reduction methods, we assume that the dimensionality of the final feature space is small compared to that of the initial feature space, that is, $K \ll D$. Our implementation is then dominated by the initial data matrix of size $(N \times D)$ and a distance matrix of size $(N \times N)$. Therefore, the memory complexity is dominated by $O(ND + N^2)$ in both methods, which is efficient when D is large.

In summary, LDA and ITML are generally fast methods when the dimensionality D is low. However, they are difficult to use when D is large due to their memory complexities. PARCA runs significantly faster than Logistic PARCA due to a faster convergence rate of the cost function with respect to the number of iterations. As both the computational complexity and the memory complexity are linear in D, PARCA is especially efficient for high-dimensional data sets. Although the slower running time of Logistic PARCA is a limitation, we have to recall that Logistic PARCA is primarily useful to uncover the latent structure of the data (examples are given in Sect. 4.4). When interpretability is not needed, PARCA should be chosen over Logistic PARCA.

4.3.5 Summary

In this section, we compared the ranking performance of PARCA and Logistic PARCA with four other distance metric learning and dimensionality reduction methods, using the Euclidean distance in the initial feature space as the baseline. The comparison was

performed using 11 data sets, showing different numbers of examples, features, classes, sparsity levels, and data types. As expected, the performance of PCA and the Euclidean distance does not depend much on the training ratio. On six out of eleven data sets (Balance, 20 Newsgroups, Reuters, WebKB, UMIST, and USPS), PCA performs noticeably worse than the Euclidean distance. On the five remaining data sets (Breast cancer, Ionosphere, Wine, Binary, and MNIST), their performances are equivalent. PCA is a lossy, unsupervised dimensionality reduction approach. There is no guarantee that the principal components will contain information relevant for data ranking. Therefore, PCA is not a good choice for the ranking task.

The performance of LDA, NCA, and PARCA depends on the value of the training ratio. LDA is especially sensitive to the amount of training data. On seven out of ten data sets (Breast cancer, 20 Newsgroups, Reuters, WebKB, Binary, MNIST, and USPS; UMIST is excluded), LDA performs worse than the Euclidean distance for low values of the training ratio. However, LDA performs well when sufficient training data are available. On five out of seven data sets (Balance, Breast cancer, Ionosphere, Wine, and USPS; UMIST and the three text data sets are excluded), LDA and PARCA are the best two methods for high values of the training ratio. In summary, LDA can achieve good ranking performance when sufficient training data are available. However, it suffers from two severe drawbacks. First, the computational cost of LDA is high when the number of dimensions is high. Therefore, it cannot be used for large-scale data sets. Second, it is prone to overfitting, leading to poor performance when only a few training data are available.

The performance of ITML does not depend much on the training ratio, which suggests that ITML needs only a few training data to achieve optimal performance. Except for the Binary alphadigits data set where it outperforms the other methods, ITML shows intermediate results. Its performance on the UMIST data set could not be assessed due to the high number of dimensions D. As we previously explained, the Mahalanobis matrix considered in ITML is of size $(D \times D)$, which makes it difficult to use with high-dimensional data sets. In addition, because ITML does not consider a projection matrix, it cannot be used for visualization purpose.

NCA shows good performance for the ranking task when a few training data are available. On nine out of eleven data sets (Balance, Ionosphere, 20 Newsgroups, Reuters, WebKB, Binary, MNIST, UMIST, and USPS), the ranking accuracy of NCA is close to the best results for low values of the training ratio. However, on two out of ten data sets (Balance and Ionosphere), the ranking performance steadily degrades as more training data become available. Although we may expect the opposite, recall that NCA learns a distance function for the classification setting rather than the ranking setting. We suspect that for some data sets, the optimization of a classification error is driving NCA away from a good distance function for ranking.

For high training ratios, PARCA shows the best ranking performance on seven out of eleven data sets (Ionosphere, 20 Newsgroups, Reuters, WebKB, MNIST, UMIST, and USPS). For low training ratios, PARCA shows the best ranking performance on eight out of eleven data sets (Balance, Breast cancer, Ionosphere, 20 Newsgroups, Reuters, WebKB, MNIST, and USPS). In addition, PARCA only needs a few training data to achieve good ranking error. For nine data sets out of eleven (Breast cancer, Ionosphere, Wine, 20 Newsgroups, Reuters, WebKB, Binary, MNIST, and USPS), PARCA performs better than Logistic PARCA. On the Balance and UMIST data sets, their prediction performance is identical.

In summary, the experiments show that PARCA offers three desirable properties. First, it is the only method that achieves good ranking performance for all values of the training ratios. Second, PARCA only needs a few training data to learn a good distance function for the

Table 3 Categories in the 20 Newsgroups data set

alt.atheism	comp.windows.x	rec.sport.hockey	soc.religion.christian
comp.graphics	misc.forsale	sci.crypt	talk.politics.guns
comp.os.ms-windows.misc	rec.autos	sci.electronics	talk.politics.mideast
comp.sys.ibm.pc.hardware	rec.motorcycles	sci.med	talk.politics.misc
comp.sys.mac.hardware	rec.sport.baseball	sci.space	talk.religion.misc

ranking task. Third, unlike LDA and ITML, PARCA is memory-efficient and can be applied to high-dimensional data sets. In addition, PARCA usually performs better than Logistic PARCA. This was expected, as Logistic PARCA is a constrained version of PARCA (using non-negativity constraints). Nonetheless, PARCA and Logistic PARCA achieve similar performance on two data sets. We will show the usefulness of such non-negativity constraints in the next section.

4.4 Interpretability

In this section, we show how Logistic PARCA is used to interpret the latent structure of the data set. We give two examples, using text data and image data.

4.4.1 Text data

We used Logistic PARCA to identify the latent topics of the 20 Newsgroups data set. This data set contains text messages posted on 20 Usenet newsgroups. The newsgroups titles are shown in Table 3. We can see that the general thematics in this data set include sport, religion, computers, etc. We removed non-alphanumerical words and common words. We also removed duplicate documents and empty documents, where a document is a text message posted on Usenet (see Sect. 4.1.2). We selected the 500 most informative words according to information gain [44]. Then, we computed the tf-idf representation of the documents in order to give high weights to important words [31]. Before we can use Logistic PARCA with the processed data set, we had to set two user-defined parameters: the number of latent topics K and the parameter β that enforces feature clustering. In order to summarize the document collection, we needed to identify a small number of latent topics where each topic is shared by several newsgroups. Therefore, K should be less than the number of newsgroups. In our experiments, we tried $K \in \{5, 10, 15\}$. As we explained in Sect. 3.3, β represents the compromise between ranking the documents and clustering the words into latent topics. Without prior knowledge, we gave equal importance to both tasks, that is, $\beta = 1$. In the following, we show the results for $K = 10$. However, similar results were obtained for $K = 5$ and $K = 15$.

In the resulting projection matrix $\mathbf{A}$, each row vector $\mathbf{A}_k$ represents a latent topic and each coefficient $\mathbf{A}_{kd}$ represents the membership strength of word d for topic k. Therefore, the topic k is defined as the set of words with the highest membership coefficients $\mathbf{A}_{kd}$. In Table 4, we show five of these clusters. We can see that in each cluster identified by Logistic PARCA, the words are strongly semantically related. The corresponding topics are mainly related to religion, computer hardware, computer software, motorized vehicles, and sport. When we compared them to the categories in Table 3, we found that each latent topic identified by Logistic PARCA is common to several categories. Hence, Logistic PARCA is able to

Table 4 Word clusters identified by Logistic PARCA in the 20 Newsgroups data set

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
islam	monitor	program	car	stats
church	card	screen	ford	hit
god	computer	xlib	engine	hockey
muslim	upgrade	window	miles	team
morality	bios	motif	article	year
clh	pc	xterm	auto	won
keith	sale	code	cars	burns
islamic	bus	widget	engines	batting
gods	scsi	sunos	rear	hes
bible	hd	issues	digital	baseball
christianity	sell	clipper	cup	scoring
question	key	mit	motorcycles	nhl
belief	house	part	police	play
sin	forsale	keys	bmw	fans
hell	controller	appreciated	dod	game
christian	port	systems	state	pens
atheists	mode	algorithm	bike	bruins
biblical	system	cryptography	manual	rangers
evil	offer	sun	ca	players
years	buying	client	sky	stanley

Each word cluster corresponds to a latent topic discussed in several newsgroups

give a quick summary of this large document collection, thereby helping users visualize and understand the structure of the collection.

4.4.2 Image data

We also used Logistic PARCA to decompose images into parts. We used the UMIST faces data set, which contains face pictures of 20 people, taken under varying angles and light levels. Each picture, represented by a (112 × 92) pixels image, is flattened into a vector of size 10304. We had to set the number of latent topics K and the parameter β that enforces feature clustering. In order to discover interesting human face features, we needed to identify a small number of latent features where each latent feature is shared by several people. Therefore, K should be less than the number of people. In our experiments, we tried $K \in \{5, 10, 15\}$. Similarly to the text data experiments and without prior knowledge, we gave equal importance to picture ranking and pixel clustering, that is, $\beta = 1$. In the following, we show results for $K = 15$. However, similar results were obtained for $K = 5$ and $K = 10$.

In the resulting projection matrix $\mathbf{A}$, each row vector $\mathbf{A}_k$ corresponds to a latent feature and each coefficient $\mathbf{A}_{kd}$ represents the membership strength of pixel d for the latent feature k. In order to visualize the latent feature, we rescaled each row vector so that the coefficients are in [0, 256] and used them to reconstruct grayscale pictures. In Fig. 10, we show six latent features identified by Logistic PARCA. These latent features seem to correspond to facial features, for example, forehead, hair, nose, mouth, ears, and cheeks. Hence, Logistic PARCA is able to decompose images into meaningful parts and can be used to isolate interesting visual entities in large image collections.

Fig. 10 Facial features identified by Logistic PARCA from the UMIST faces data set

5 Conclusion

In this work, we studied the distance metric learning problem in the ranking framework, rather than in the conventional classification framework. First, we defined a ranking error between vectors and considered a linear model, that is, a Mahalanobis distance. Then, we proposed PARCA, our distance metric learning method, to optimize this ranking error. In order to apply PARCA to large-scale data sets, we also proposed an efficient implementation of our approach. Then, we evaluated the performance of PARCA on 11 heterogeneous data sets. These data sets differ by the numbers of examples, dimensions, and classes, the sparsity levels, and the data types.

We compared PARCA and Logistic PARCA with four other distance metric learning and dimensionality reduction methods: principal component analysis, linear discriminant analysis, neighborhood components analysis, and information-theoretic metric learning. We also used the Euclidean distance in the initial feature space as the baseline. The results showed that PARCA achieves the best ranking errors for most data sets and most values of the training ratios. We also showed that unlike LDA, PARCA only needs a few training data to achieve good ranking error. This property is important for applications where labeled examples are costly and difficult to obtain. In addition, our approach is simple to implement and efficient and can be used for better visualization and understanding of the data set.

Because we think interpretability is an important property of a distance metric learning approach, we also presented an extension of PARCA, called Logistic PARCA. Using the logistic function, we applied positivity constraints on the projection matrix. Then, we defined

an entropy-based cost function that forces the model to cluster the original features into meta-features. As a result, Logistic PARCA is able to uncover the latent structure of the data, while having the same computational complexity as standard PARCA. We tested the interpretability property of Logistic PARCA on two data sets and showed that it is able to identify the latent topics discussed in the 20 Newsgroups data set and to decompose the UMIST faces into meaningful parts. In the case of large data sets, Logistic PARCA can help people quickly visualize, summarize, and better understand the data.

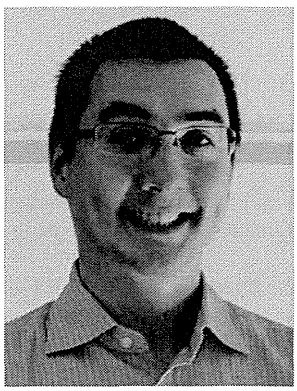
Although we obtained good experimental results, PARCA and Logistic PARCA suffer from two limitations. First, they are essentially linear methods. Previous studies have suggested learning non-linear distance functions using the kernel trick [18]. In our future work, we wish to extend our approach to non-linear distance metric learning. The second limitation is related to the cost function optimized by PARCA and Logistic PARCA. We defined a ranking error that uniformly penalizes all the pairs incorrectly ordered for a given list. With this ranking error, a wrong prediction of the highest ranked documents has the same cost as a wrong prediction of the lowest ranked documents. However, in the information retrieval setting, we are mainly interested in the results at the top of the list and the quality prediction in the bottom of the list is of little importance. In our future work, we wish to extend our approach to ranking errors that focus on the ranking accuracy of higher ranked documents [6,39], while still providing interpretability properties.

References

1. Amini MR, Truong TV, Goutte C (2008) A boosting algorithm for learning bipartite ranking functions with partially labeled data, In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, USA, pp 99–106
2. Baccini A, Dejean S, Lafage L et al (2011) How many performance measures to evaluate information retrieval systems? Knowl Inform Syst 30(3):693–713
3. Baker LD, McCallum AK (1998) Distributional clustering of words for text classification, In: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, USA, pp 96–103
4. Bekkerman R, El-Yaniv R, Tishby N et al (2003) Distributional word clusters vs. words for text categorization. J Mach Learn Res 3:1183–1208
5. Burges S, Shaked T, Renshaw E et al (2005) Learning to rank using gradient descent. In: Proceedings of the 22nd international conference on machine learning. ACM, New York, NY, USA, pp 89–96
6. Burges CJC, Ragno R, Le QV (2007) Learning to rank with nonsmooth cost functions. In: Advances in neural information processing systems, vol 19. MIT Press, pp 193–200
7. Chapelle O, Shivaswamy P, Vadrevu S et al (2010) Multi-task learning for boosting with application to web search ranking. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, NY, USA, pp 1189–1198
8. Chen Y, Rege M, Dong M et al (2008) Non-negative matrix factorization for semi-supervised data clustering. Knowl Inform Syst 17(3):355–379
9. Cohen WW, Schapire RE, Singer Y (1999) Learning to order things. J Artif Intell Res 10(1):243–270
10. Cover TM, Thomas JA (1991) Elements of information theory. Wiley, London
11. Davis JV, Kulis B, Jain P et al (2007) Information-theoretic metric learning. In: Proceedings of the 24th international conference on machine learning. ACM, New York, NY, USA, pp 209–216
12. Dela Rosa K, Metsis V, Athitsos V (2011) Boosted ranking models: a unifying framework for ranking predictions. Knowl Inform Syst 30(3):543–568
13. Dhillon IS, Modha DS (2001) Concept decompositions for large sparse text data using clustering. Mach Learn 42(1):143–175
14. Duda RO, Hart PE, Stork DG (2000) Pattern classification. Wiley, London
15. Forman G (2003) An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 3:1289–1305
16. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Proceedings of the 13th international conference on machine learning. ACM, New York, NY, USA, pp 148–156

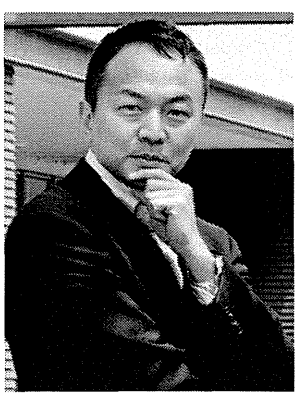
17. Freund Y, Iyer R, Schapire RE et al (2003) An efficient boosting algorithm for combining preferences. J Mach Learn Res 4:933–969
18. Globerson A, Roweis S (2006) Metric learning by collapsing classes. In: Advances in neural information processing systems, vol 19. MIT Press, pp 451–458
19. Goldberger J, Roweis S, Hinton G et al (2004) Neighbourhood components analysis. In: Advances in neural information processing systems, vol 17. MIT Press, pp 513–520
20. Harpeled S, Roth D, Zimak D (2003) Constraint classification for multiclass classification and ranking. In: Advances in neural information processing systems, vol 16. MIT Press, pp 785–792
21. Huang K, Ying Y, Campbell C (2011) Generalized sparse metric learning with relative comparisons. Knowl Inform Syst 28(1):25–45
22. Jain P, Kulis B, Dhillon IS et al (2008) Online metric learning and fast similarity search. In: Advances in neural information processing systems, vol 21. MIT Press, pp 761–768
23. Jolliffe I (1986) Principal component analysis. Springer, New York
24. Kulis B, Sustik M, Dhillon IS (2006) Learning low-rank kernel matrices. In: Proceedings of the 23rd international conference on machine learning. ACM, New York, NY, USA, pp 505–512
25. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791
26. Lee DD, Seung HS (2001) Algorithms for Non-negative Matrix Factorization. In: Advances in neural information processing systems. MIT Press, pp 556–562
27. Liu TY (2009) Learning to rank for information retrieval. Found Trends Inform Retriev 3(3):225–331
28. Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, New York
29. Martínez AM, Kak AC (2001) PCA versus LDA. IEEE Trans Pattern Anal Mach Intell 23(2):228–233
30. Pereira F, Tishby N, Lee L (1993) Distributional clustering of English words. In: Proceedings of the 31st annual meeting on association for computational linguistics. ACL, Stroudsburg, PA, USA, pp 183–190
31. Salton G, McGill MJ (1986) Introduction to modern information retrieval. McGraw-Hill, Inc., New York
32. Schultz M, Joachims T (2004) Learning a distance metric from relative comparisons. In: Advances in neural information processing systems, vol 16. MIT Press, pp 41–48
33. Shalev-Shwartz S, Singer Y, Ng AY (2004) Online and batch learning of pseudo-metrics. In: Proceedings of the 21st international conference on machine learning. ACM, New York, NY, USA, pp 743–750
34. Shental N, Hertz T, Weinshall D et al (2002) Adjustment learning and relevant component analysis. In: Proceedings of the 7th European conference on computer vision. Springer, London, UK, pp 776–792
35. Slonim N, Tishby N (2000) Document clustering using word clusters via the information bottleneck method. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, USA, pp 208–215
36. Sugiyama M (2006) Local Fisher discriminant analysis for supervised dimensionality reduction. In: Proceedings of the 23rd international conference on machine learning. ACM, New York, NY, USA, pp 905–912
37. Thurau C, Kersting K, Wahabzada MC et al (2010) Convex non-negative matrix factorization for massive datasets. Knowl Inform Syst 29(2):457–478
38. Usunier N, Amini MR, Gallinari P (2005) Generalisation error bounds for classifiers trained with interdependent data. In: Advances in neural information processing systems, vol 18. MIT Press, pp 1369–1376
39. Usunier N, Buffoni D, Gallinari P (2009) Ranking with ordered weighted pairwise classification. In: Proceedings of the 26th international conference on machine learning. ACM, New York, NY, USA, pp 1057–1064
40. Wang D, Li T, Ding C (2010) Weighted feature subset non-negative matrix factorization and its applications to document understanding. In: Proceedings of the 2010 IEEE international conference on data mining, pp 541–550
41. Weinberger KQ, Blitzer J, Saul LK (2006) Distance metric learning for large margin nearest neighbor classification. In: Advances in neural information processing systems, vol 18. MIT Press, pp 1473–1480
42. Xing EP, Ng AY, Jordan MI et al (2002) Distance metric learning, with application to clustering with side-information. In: Advances in neural information processing systems, vol 15. MIT Press, pp 505–512
43. Xu W, Liu X, Gong Y (2003) Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, USA, pp 267–273
44. Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. In: Proceedings of the fourteenth international conference on machine learning. Morgan Kaufmann Publishers, San Francisco, USA, pp 412–420

Author Biographies



Jean-François Pessiot is a postdoctoral researcher at the National Institute for Advanced Industrial Science and Technology in Tokyo, Japan. He received his PhD in Computer Science from the Pierre-and-Marie-Curie University in 2008. His research interests include data mining, machine learning, and computational biology.

Hyeryung Kim is currently a research scientist at Dong-A Pharmaceutical in Seoul, South-Korea. She received her PhD in Bioinformatics from the Tokyo Medical and Dental University. Her research interests include kernel canonical correlation analysis, analysis of comprehensive gene expression profiles, and study of drug effects on human cells.



Wataru Fujibuchi received his PhD degree from the Department of Biophysics, Kyoto University, in 1998. From 1999 to 2003, he worked as a Visiting Fellow and Staff Scientist at the NCBI, USA. From 2003, he worked as a Research Scientist of the National Institute of Advanced Industrial Science, Japan, and in 2007, he became a Team Leader. Now, he is a Professor at the Center for iPS Research and Application, Kyoto University. His research interests include cell state-space analysis, large-scale gene expression data mining, microarray standardization, speed-up of biocalculation by accelerator, and systems biology.

Inference of Gene Regulatory Networks to Detect Toxicity-Specific Effects in Human Embryonic Stem Cells

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

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

Reiko Nagano, Hideko Sone
Research Center for Environmental Risk
National Institute for Environmental Studies
Tsukuba, Japan
nagano.reiko@gmail.com
hsone@nies.go.jp

Satoshi Imanishi, Seiichiroh Ohsako
Center for Disease Biology and Integrative Medicine,
Graduate School of Medicine, The University of Tokyo
Tokyo, Japan
imanishi@m.u-tokyo.ac.jp
ohsako@m.u-tokyo.ac.jp

***Abstract*—Environmental chemicals are known to cause serious developmental problems in embryos. To prevent injurious chemical effects, knowledge of the chemical toxicity mechanisms in human embryos is important. To reveal the functional mechanisms in living cells, inferring a gene regulatory network is a useful approach. We applied our developed statistical methods based on Structural Equation Modeling to infer the gene regulatory networks in human embryonic stem cells. In this study, we improved the SEM approach and applied this enhanced version to expression profiles in human embryonic stem cells exposed to various chemicals. For almost all of the tested chemicals, the cell differentiation-related genes and the neuron development-related genes were intermixed in the inferred networks. Since the chemicals' networks displayed diffusion type shapes, the effects of chemical toxicity are considered to affect a few target genes at first, and then ultimately many genes via regulatory mechanisms. Furthermore, the genes that were finally affected were conserved among chemicals with the same toxicity: Tuj1 in Neurotoxic chemicals, Oct3/4 and Pax6 in Genotoxic chemicals, and Oct3/4 in Carcinogenic chemicals. These finally affected genes are considered to be the results of toxicity-specific effects in ES cells, and they reflected the features of the toxicity. We also found that some chemicals shared the same regulatory mechanism. The detected toxicity-specific effects are valuable for developing methods to prevent chemicals from disturbing normal development.**

Keywords-Structural Equation Modeling; Gene Regulatory Network; Embryonic Stem Cell; Environmental Chemicals

I. INTRODUCTION

We are exposed to many chemicals, which are produced by our usual life activities. Since the toxicity of environmental chemicals is known as one of the typical factors causing developmental toxicity, we investigate the specific effects of chemical toxicity [1]. 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 [2][3]. For example, Methylmercury is known as a developmental toxin that affects fetal development [4][5]. Furthermore, certain chemicals can cause serious developmental problems and abnormal cell differentiation in embryos [6][7][8].

To prevent the harmful effects of chemicals, elucidation of the toxic stress response in embryonic cells is crucial [9][10]. A gene regulatory network is a useful approach to reveal the regulatory mechanisms in living cells. Using 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 [11][12]. In our previous 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 [13][14]. 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, due to insufficient information in the gene expression profiles. To clarify the mechanisms of biological processes in living cells, un-observed factors that 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) in combination with factor analysis and a four-step procedure [15][16]. This approach allowed us to reconstruct a model of transcriptional regulation that involves protein-DNA interactions from only the gene expression data, in the absence of protein

2013, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

information [15]. The significant features of SEM are the inclusion of latent variables within the constructed model and the ability to infer the network, including its cyclic structure. Furthermore, the SEM approach allows us to strictly evaluate the inferred model by using fitting scores. The SEM approach is useful for detecting the causality among selected genes, as the linear relationships between genes are assumed to minimize the difference between the model's covariance matrix and the calculated sample covariance matrix [17][18][19]. Some fitting indices are defined for evaluating the model adaptability, and thus the most suitable model can be selected by SEM [1][19].

Here, we applied the SEM approach to infer the regulatory network among 9 development-related genes. The mRNA levels of these 9 genes were measured in human embryonic stem cells exposed to 15 environmental chemicals. The chemicals were considered to have developmental toxicities that adversely affect the developmental process in human embryos. Thus, inferring the gene regulatory network among development-related genes will help us to elucidate the toxic stress response in the human embryo. Furthermore, we improved our SEM approach for constructing preliminary initial models from the time-series data, in the absence of known regulatory interactions among the genes. We applied this improved SEM approach to infer the chemical-specific regulatory network among the development-related genes.

II. MATERIALS AND METHODS

A. Expression data

We utilized expression data that were measured to clarify the effects of chemical toxicity on neuronal differentiation [7][20]. In these expression data, nine genes considered to be affected by chemicals were measured in human embryonic stem cells: GATA2, Nanog, Oct3/4, Nodal, Lmx1A, MAP2, Nestin, Pax6, and Tuj1 [7][20]. Among the 9 genes, GATA2, Nanog, Oct3/4, and Nodal are mainly related to cell differentiation, and the other genes are related to neuron development. As an internal control, the expression of beta-actin was also measured. The expression data of these 10 genes were obtained from human embryonic stem cells exposed to 15 chemicals: Methylmercury (MeHg), 2-Nitropropane (2-NP), Acrylamide (ACA), p-Nitroaniline (p-NA), 4-hydroxy PCB107 (PCB), Benzo[a] pyrene (BZP), Diethylnitrosamine (DENA), Diethylaminofluorene (DEAF), Phenobarbital (PB), Tamoxifen (TMX), Diethylstilbestrol (DES), TCDD (TCDD), Thalidomide (THAL), Bisphenol-A (BPA), and Permethrin (PER) [7][20]. The toxicity of each chemical was classified into one of four types: Neurotoxic (MeHg, 2-NP, ACA, p-NA, and PCB), Genotoxic (BZP, DENA, and DEAF), Carcinogenic (PB, TMX, DES, and TCDD), and others (THAL, BPA, and PER). 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 genes was measured twice under each condition by RT-PCR, and thus 600 (15 chemicals x 4 time periods x 5 concentration types x 2 repeats) expression patterns per gene were measured [20].

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. 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. Multi-factor analysis of variance

In this study, the data contained three factors that affect gene expression: chemicals, exposure times, and concentrations. To detect the significant factors for differences in gene expression, we applied the analysis of variance (ANOVA) for multiple factors [21]. Although the multi-factor ANOVA model includes each factor's effect and all combinations of interactions between the factors, the triple interactions among the factors were confounded with error terms, because the data lacked repetition [21]. Therefore, we used the linear effects model for analysis:

$$E_{ijk}=\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}+\varepsilon_{ijk} \quad (2)$$

where E_{ijk} is the expression level of each gene under one condition, μ is the averaged value of all measured data, α_i is the effect of factor A, β_j is the effect of factor B, γ_k is the effect of factor C, $(\alpha\beta)_{ij}$ is the interaction between factors A and B, and ε_{ijk} is the error term.

Depending on the linear effects model, the total some of squares, S_{Total} could be decomposed into the following components:

$$S_{Total}=S_A+S_B+S_C+S_{AB}+S_{AC}+S_{BC}+S_e \quad (3)$$

where S_A, S_B, and S_C mean the sum of the squared differences between each factor's marginal mean and the overall mean; S_{AB}, S_{AC}, and S_{BC} mean the sum of the squared differences for particular corresponding data means, marginal means, and overall mean; and S_e measures the difference between S_{Total} and the total sum of squares of all effects. The degree of freedom for S_{Total} was the number of all observed data minus one, and the degrees of freedom for S_A, S_B, and S_C were the number of levels for the factor minus one. The mean square values for S_A, S_B, and S_C were the sums of the squares divided by the numbers of degrees of freedom. In S_e, the degree of freedom was the total degrees of freedom minus the sum of the factor degrees of freedom. The mean square of S_e was the sum of the squares divided by the number of degrees of freedom. In the analysis of variance, S_{Total} accounted for the

2013, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

factor effects (S_A, S_B, S_C, S_{AB}, S_{AC}, S_{BC}) and the contribution of S_e.

To compare the factor effects, the statistical F-test was used. The F statistic is the mean square for the factor divided by the mean square of the error terms. This F statistic is known to follow an F distribution with degrees of freedom for each factor effect and degrees of freedom for the error terms. Thus, we could calculate the probabilities of the factor effects from the F statistics.

C. *Extraction of causalities from expression data*

In an SEM analysis, an initial model should be assumed, but no regulations were defined among the selected genes in this study. Thus, we had to construct an initial model among the 9 genes for each chemical. To detect the regulatory relationships between the gene pairs from the measured time series expression data, we applied cross correlation coefficients 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 [22]. The cross correlation values ranged between −1 and +1. In a time series analysis, the cross correlation between two time series describes the normalized cross covariance function. Let $X_t = \{x_1, \ldots, x_N\}$, $Y_t = \{y_1, \ldots, y_N\}$ represent two time series data including N time points. The cross correlation is then given by

$$r_{xy} = \frac{\sum_{t=1}^{N}(x_i - \bar{x})(y_{i+d} - \bar{y})}{\sqrt{\sum_{t=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{t=1}^{N}(y_{i+d} - \bar{y})^2}} \quad (4)$$

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 $d = -1, 0, +1$.

D. *Construction of the initial model*

In this study, we inferred the chemical-specific regulatory network, and thus the differences between times and concentrations could be merged for the construction of the initial model. Fig. 1 shows the new method developed for constructing an initial model of each chemical, with the merging of several conditions. First, we constructed lag matrices to merge the time difference. The time difference was summarized by the time lag values in the cross correlations among genes. Since the time lags indicated the order of the expression pattern among the gene pairs, the rough causality between all gene pairs could be extracted. In this study, three cross correlations were calculated with three lags, −1, 0, and +1, and the three absolute values of the cross correlations were compared. The value d with the highest absolute value was selected as the causal information between the gene pairs, and the selected lag value d wasarranged as a matrix element in a lag matrix.

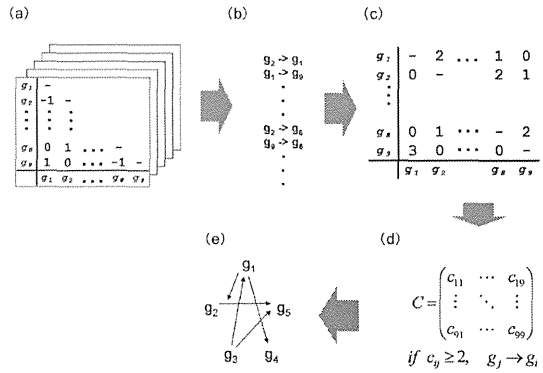


Figure 1. Procedure for initial model construction: (a) Time-lag matrices for each chemical. Five time-lag matrices were obtained for each chemical. (b) Binomial relationships. (c) Frequency matrix of causal relationships between all gene pairs. (d) Selection of possible causal relationships from the frequency matrix. (e) Construction of an initial model with selected causal relationships.

Lag Matrices were constructed for each concentration of a chemical. Thus, five time lag matrices were constructed for each chemical (Fig. 1a).

We subsequently merged the concentration difference of each chemical. For each chemical, there are five lag matrices according to the chemical concentrations, and we considered that the chemical-specific relationships among the genes would be conserved in several lag matrices. To obtain the chemical-specific relationships among the genes, we extracted the binary relationships between gene pairs from the five lag matrices for each chemical. If the same relationships existed in several lag matrices, then the binomial relationships were duplicated (Fig. 1b).

In the next step, we constructed one frequency matrix for each chemical. From the binary relationships, we counted the frequencies of all gene regulatory pairs, and each frequency number was arranged as an element of a frequency matrix (Fig. 1c). In this step, the concentration difference could be merged, since the elements of the frequency matrix indicate the information for the different concentrations. We subsequently selected the gene pairs with frequency matrix values greater than or equal to two, as the chemical-specific regulation (Fig. 1d). At the final step, we constructed an initial model for each chemical from the extracted relationships between the genes (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. By using this approach, an initial model can include cyclic structures.

E. *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. In general, SEM is a comprehensive statistical model that includes two types of variables: observed and latent. These

2013, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

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 9 genes (GATA2, Nanog, Oct3/4, Nodal, Lmx1A, MAP2, Nestin, Pax6, and Tuj1) were defined as the observed variables. Meanwhile, none were defined as latent variables, since considerations about the common regulator of several genes are dispensable for this study. The un-observed factor, which affected each gene's expression, was calculated as an error. 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:

$$y=\Lambda y+e \quad (5)$$

Here, y is a vector of p observed variables (measured gene expression patterns), and Λ is a p x p matrix representing the regulatory relationships between the observed variables. Errors that affect the observed endogenous variables are denoted by e.

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 Chi-square statistic (CMIN), 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 were considered to be useful to clarify the degree of model fitting in this study.

F. Parameter estimation

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

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

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 [23].

G. Iteration for the 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 the measured data can be selected.

In the network model, the covariance matrix between variables is calculated by the estimated parameters. The similarity between the 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 the quantitative similarity between a constructed model and an actual relationship, fitting scores are usually utilized. In this study, the quality of the fit was predicted by four different fitting scores: CMIN(Prob), GFI, AGFI, CFI, and RMSEA. The value of CMIN(Prob) is calculated by the Chi-square statistic divided by the degrees of freedom, and a CMIN(Prob) value higher than 0.05 is considered as a good model fit. 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 [24]. 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:

Step 1: Deletion of a non-significant edge from the 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 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 is also 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 [24]. The MI value indicates the

2013, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

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. 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.

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, 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 within the network.

H. Extraction of association rules by affinity analysis

We applied affinity analysis to discover the similar regulatory mechanism models among the 15 chemicals' networks. To detect the relative chemical pairs as association rules, we created a binary dataset with conserved regulations among different chemicals. According to the original definition of association rule mining [25], we defined the problem of association rule mining as follows: Let $I = \{ i_1,\ldots, i_n \}$ be a set of n binary attributes called items. Let $T = \{ t_1,\ldots, t_m \}$ be a set of database transactions. Each transaction t_k is represented by the binary vector $t_k = (t_k^1, t_k^2, \ldots, t_k^n)$, which includes n elements. The value of t_k^l indicates the appearance of transaction t_k in item i_l. In this study, the 15 chemicals were defined as a set of items, and each conserved gene regulation between the different chemicals was considered as one transaction. Thus, the value of 1 indicated the appearance of the conserved gene regulation in the chemical's network, while the value of 0 indicated its absence.

An association rule is defined as the implication of the form $I_a \Rightarrow I_b$, where I_a and I_b are sets of some items in I, but some of the same items are not present in I_a and I_b. To detect the association rules, we used some constraints: support, confidence and lift. Support is defined as the proportion of transactions that contain the item set to all transactions. Thus, $support(I_a,I_b)=prob(I_a,I_b)$ was calculated as the joint probability of I_a and I_b. The confidence constraint is displayed as $conf(I_a \Rightarrow I_b)$, and it is defined as the conditional probability $prob(I_a|I_b)$. Thus, we calculated $conf(I_a \Rightarrow I_b)$ from the proportion of transactions with the item set I_b to the transactions with the item set I_a. The lift constraint is defined as:

$$lift(I_a \Rightarrow I_b) = conf(I_a \Rightarrow I_b)/prob(I_b) \quad (7)$$

Lift is a measure of the performance of an association rule with respect to the population as a whole, against the random choice. Thus, lift was obtained by calculating the ratio of the target response to the average response. In general, a lift value over 1 is suitable for association rules.

III. RESULTS AND DISCUSSION

A. Chemical concentrations had no effect

In this study, gene expression was measured in the presence of different concentrations of various chemicals, with several exposure times. To reveal the most effective factor for gene expression, multi-factor ANOVA was applied to the measured data. In statistics, ANOVA is utilized to detect differences between groups in terms of some variables. Usually, the chance of committing a type I error will increase by performing multiple two-sample t-tests, and a statistical test is needed to determine whether or not the means of more than two groups should be applied, such as Tukey's HSD test and so on. Although these post-hoc tests are useful for detecting the factor pairs with significant differences between them, the factor pairs are not important in this study. Instead, we wanted to determine factors, which caused gene expression differences, and thus we compared three factors: chemicals, time differences, and concentrations.

The 15 chemicals were divided into 3 categories by their toxicities: Neurotoxic chemicals, Genotoxic chemicals, Carcinogenic chemicals, and other type chemicals. We compared the gene expression differences between these toxicity types. We calculated a p-value from the F statistic for each gene. The p-value is the probability that the variation between conditions may have occurred by chance, so genes with smaller p-values vary more significantly. Thus, the gene's variation is less likely to have occurred by chance, and is conversely more likely to be connected to the difference in conditions. The probabilities of expression differences for each gene, grouped by each factor, are shown in Table I. Interestingly, the expression of all of the genes was significantly different among the chemicals and the time differences. However, the chemical concentrations showed almost no significant differences in terms of the expression of the genes. Thus, the concentrations of the chemicals had no effect on the expression of the tested genes in the ES cells.

TABLE I. RESULTS OF MULTI-FACTOR ANOVA

	Chemical (a)	Concentration (b)	Time (c)	a * b	a* c	b * c
GATA2	<0.01	0.076	<0.01	0.559	<0.01	0.450
Nanog	<0.01	<0.01	<0.01	0.011	<0.01	0.022
Oct34	<0.01	<0.01	<0.01	0.055	<0.01	0.044
Nodal	<0.01	0.130	<0.01	<0.01	<0.01	0.040
Lmx1A	<0.01	0.714	<0.01	<0.01	<0.01	0.787
MAP2	<0.01	0.479	<0.01	<0.01	<0.01	0.576
Nestin	<0.01	<0.01	<0.01	0.012	<0.01	0.548
Pax6	<0.01	0.575	<0.01	<0.01	<0.01	0.861
Tuj1	<0.01	0.011	<0.01	0.810	<0.01	0.097

a. Probabilities were calculated from the F statistics and the degrees of freedom.

b. Significant probabilities are displayed as "<0.01" in this table.

2013, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

B. The complexities of the initial models are related to the chemical toxicity

We utilized our newly developed method to construct the initial gene regulatory network models under the conditions with 15 chemicals. One of the distinguishing features of our new method is its ability to include a cyclic structure in the network model. Cyclic regulation, such as feedback regulation, is considered to be important for cells to control normal gene expression, and the new method is useful to detect cyclic regulation from the gene expression data. Fig. 2 shows the constructed initial network models.

In Fig. 2, the components of the constructed models were: 9 genes with 22 relationships in MeHg, 9 genes with 23 relationships in 2-NP, 9 genes with 19 relationships in ACA, 9 genes with 23 relationships in p-NA, 9 genes with17 relationships in PCB, 9 genes with 9 relationships in BZP, 8 genes with 14 relationships in DENA, 8 genes with 10 relationships in DEAF, 8 genes with 19 relationships in PB, 9 genes with 23 relationships in TMX, 7 genes with 9 relationships in DES, 9 genes with 23 relationships in TCDD, 8 genes with 10 relationships in THAL, 6 genes with 9 relationships in BPA, and 8 genes with 10 relationships in PER. The distribution of the number of relationships according to the toxicity type is displayed in Fig. 3. In Figs. 2 and 3, the numbers of edges were obviously different, according to the chemicals' toxicity. Neurotoxic and Carcinogenic chemicals contained more relationships than

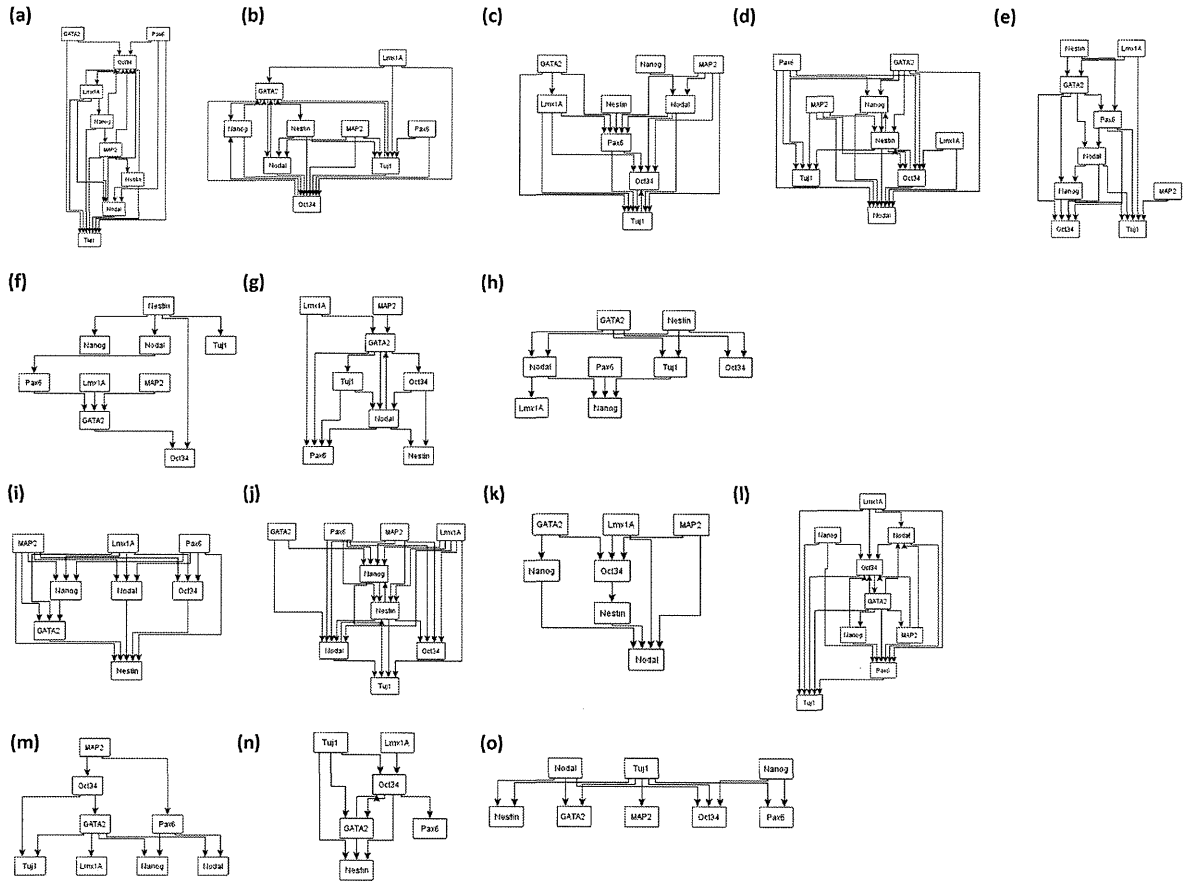


Figure 2. Initial network models: (a) MeHg, (b) 2-Np, (c) ACA, (d) p-NA, (e) PCB, (f) BZP, (g) DENA, (h) DEAF, (i) PB, (j) TMX, (k) DES, (l) TCDD, (m) THAL, (n) BPA, (o) PER. The networks with the same toxicity are arranged on the same line.

2013, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

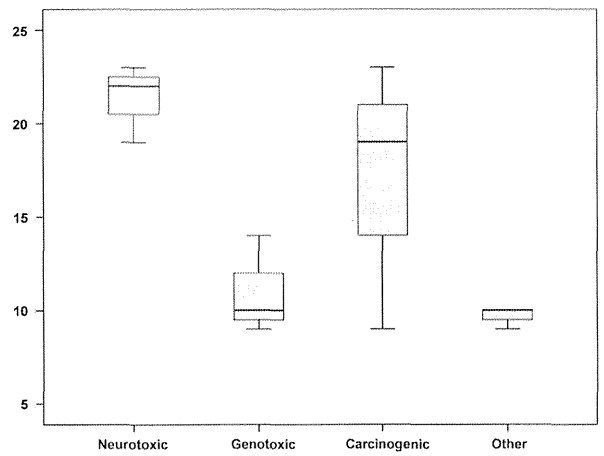


Figure 3. Box plot of edge numbers: Distribution of the number of edges in each initial model.

Genotoxic and other chemicals. Furthermore, only one or two genes were arranged as the last endogenous genes in the initial models with Neurotoxic and Carcinogenic chemicals, as opposed to two or more genes in the initial models of Genotoxic and other chemicals. Thus, the effects of the Neurotoxic or Carcinogenic chemicals were complicated, but could be summarized into only one or two target genes. In contrast, the expressions of many genes were finally affected by Genotoxic and other chemicals, via simple regulatory networks. These differences between chemical toxicity types 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. Before the SEM calculation, we simplified all of the initial models. To simplify these duplicated interactions, we only retained the longest path between two genes. In the initial model, the edges do not represent the direct regulation, but the time provenience information. In other words, the difference between direct and indirect interactions in the initial model is not very important. Thus, the regulation displayed by a direct path could be replaced by indirect paths in the model. By retaining the longest paths, all of the preceding information was included, as the simplest diagram.

C. *Structures of inferred networks*

The final inferred networks for each chemical are depicted in Fig. 4, and the goodness of fit scores are displayed in Table II. From Table II, almost all of the models were considered to fit well with the measured data by some fitting scores, CMIN(Prob), CFI, and RMSEA, except for the DES network. In the DES network, all of the fitting scores indicated that the inferred network could not be judged as a well-fitted model. Since the obtained fitting scores were the best scores in this analysis, we considered the network inference for DES to need more expression data.

The inferred networks of chemicals revealed distinct structures. The cell differentiation-related genes and the neuron development-related genes were intermixed in almost all of the inferred networks, except for MeHg and BPA. In the inferred network of MeHg, the regulations among cell differentiation-related genes and the regulation among neuron development-related genes were separated to the right and left. This specific shape means that the effects of MeHg appeared differently between neuronal and other development. This difference may be related to the two different effects of MeHg: developmental deficits in children [26], and risk of cardiovascular disease in adults [27]. On the other hand, cell differentiation-related genes and neuron development-related genes were separated at the top and bottom in the BPA network. In the BPA network, neuron development-related genes were only disturbed by cell differentiation-related genes.

TABLE II. FITTING SCORES OF INFERRED NETWORKS

	Neurotoxic					Genotoxic			Carcinogenic				Other		
	MeHg	2-NP	ACA	p-NA	PCB	BZP	DENA	DEAF	PB	TMX	DES	TCDD	THAL	BPA	PER
CMIN (*Prob*)	0.50	0.34	0.06	0.26	0.30	0.44	0.16	0.11	0.01	0.27	0.00	0.63	0.31	0.52	0.11
GFI	0.76	0.82	0.83	0.78	0.79	0.79	0.84	0.77	0.75	0.81	0.74	0.83	0.83	0.78	0.78
AGFI	0.60	0.63	0.61	0.59	0.61	0.62	0.65	0.60	0.54	0.61	0.52	0.64	0.60	0.64	0.56
CFI	1.00	0.99	0.96	0.97	0.98	1.00	0.97	0.94	0.90	0.98	0.88	1.00	0.99	1.00	0.96
RMSEA	0.00	0.07	0.15	0.10	0.08	0.03	0.12	0.14	0.21	0.09	0.23	0.00	0.08	0.00	0.14

a. Five fitting scores were utilized for measuring the fitness level between the constructed model and the measured data.

b. The well-fitted threshold of each score is: CMIN(Prob) is P>0.05, GFI > 0.90, AGFI> 0.90, CFI>0.90, RMSEA< 0.05.

2013, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

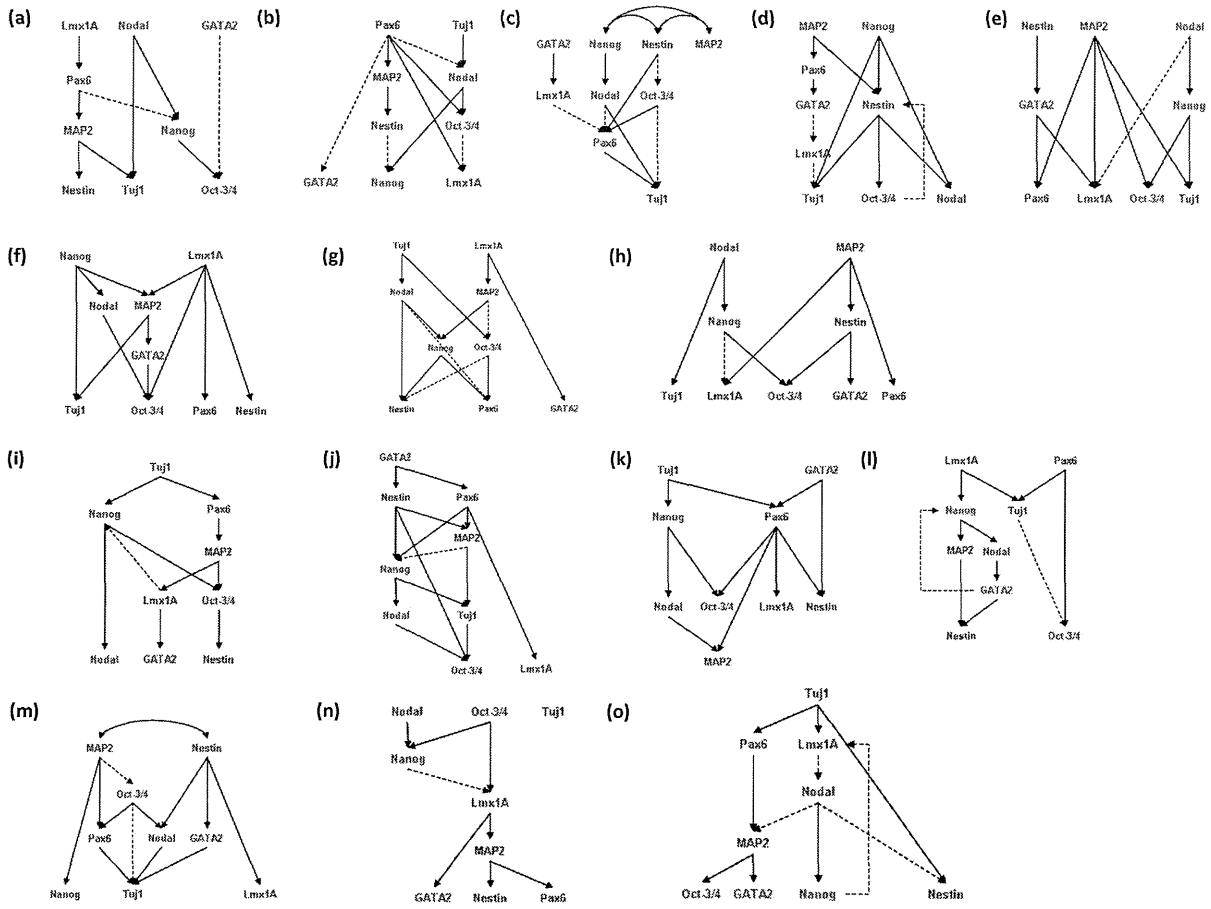


Figure 4. Inferred chemical networks: A positive relationship between genes is displayed with a solid arrow. A negative relationship between genes is displayed with a dashed arrow. Gene names with blue characters indicate "neuron development-related genes", and genes with red characters indicate "cell differentiation-related genes". (a) MeHg, (b) 2-Np, (c) ACA, (d) p-NA, (e) PCB, (f) BZP, (g) DENA, (h) DEAF, (i) PB, (j) TMX, (k) DES, (l) TCDD, (m) THAL, (n) BPA, (o) PER. The networks with the same toxicity are arranged on the same line.

Concerning the shapes of the inferred networks, we defined the network shape by comparing the numbers of genes at the top phase (*N*(*top*)) and the final phase (*N*(*bottom*)) within each chemical network. One of the specific shapes was a centralized model, which was defined as N(*top*)-N(*bottom*)≥2. In this model, many genes were arranged at the top phase, and only a few genes were arranged at the final phase in the network structure. The ACA network was the only network with a centralized model. The other specific shape was a diffusion model. The shape of a diffusion model is defined as N(*bottom*)-N(*top*)≥2. Among the well-fitted models, four networks were classified into diffusion models: BZP, DEAF, PB, and PER. The shape of the BPA network was different from those of the other networks, and resembled a bow-tie like model.

Fundamentally, the genes were hierarchically controlled in the inferred networks, but there were a few recursive relationships. Interestingly, the values of the regression weights of the recursive regulations among all of the inferred networks were negative: regulation from Oct3/4 to Nestin in the p-Na network, regulation from GATA2 to Nanog in TCDD, and regulation from Nanog to Lmx1A in PER. These recursive regulations indicated that feedback regulation exists in ES cells.

D. Detection of Toxicity-Specific Effects

To detect the specific features that were dependent on the toxicity type, we monitored the position of each gene in the inferred networks. Table III displays the number and probability of incoming edges and those of outgoing edges for each gene. Among the Neurotoxic chemicals' networks, Tuj1 has significantly few incoming edges and significantly many outgoing edges. Actually, Tuj1 was arranged as a result of network regulation in almost all of the Neurotoxic networks.

2013, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE III. INTERACTING EDGES OF EACH GENE

	Neurotoxic				Genotoxic				Carcinogenic			
	OUTPUT		INPUT		OUTPUT		INPUT		OUTPUT		INPUT	
	Num	*P*	Num	*P*	Num	*P*	Num	*P*	Num	*P*	Num	*P*
Oct3/4	4	0.113	8	0.111	2	0.120	7*	0.032	1*	0.043	9*	0.042
GATA2	5	0.135	3	0.083	1	0.091	3	0.222	6	0.146	2	0.070
Lmx1A	3	0.084	7	0.141	6	0.101	2	0.145	4	0.141	3	0.114
MAP2	9	0.077	2	0.049	7	0.070	3	0.222	5	0.155	6	0.160
Nanog	7	0.132	5	0.146	7	0.070	3	0.222	7	0.118	8	0.079
Newtin	7	0.132	5	0.146	2	0.120	5	0.174	3	0.111	6	0.160
Nodal	8	0.107	5	0.146	6	0.101	2	0.145	4	0.141	4	0.154
Pax6	9	0.077	8	0.111	0	0.060	5	0.174	10*	0.025	4	0.154
Tuj1	1*	0.031	10*	0.043	2	0.120	3	0.222	6	0.146	4	0.154

a. The significant values ($P<0.05$) are highlighted with *.

This means that the toxicities of Neurotoxic chemicals are considered to finally affect Tuj1, which is known to contribute to microtubule stability in neuronal cells [28]. Although the expression levels of 5 genes were measured as neuron development-related genes, Tuj1 was detected as the final target of Neurotoxicity.

Among the Genotoxic chemicals' networks, Oct3/4 exhibited a significant number of incoming edges. Furthermore, both Pax6 and Oct3/4 were arranged at the lower phase in all Genotoxic networks. Oct3/4 is one of the key regulators of pluripotency [29], and Pax6 is known as a key transcription factor for the development of the cerebral cortex and other sensory organs [30]. Considering the features of both Pax6 and Oct3/4, developmental processes, such as normal cell differentiation, were disturbed by Genotoxic chemicals.

In the Carcinogenic networks, both the incoming and outgoing edges of Oct3/4 were significant, and Oct3/4 was arranged as a result in almost all of the Carcinogenic networks. The chemicals that were classified as either Genotoxic or Carcinogenic are known as carcinogens [31][32]. Thus, the Genotoxic and Carcinogenic features indicated that the chemical disturbance of Oct3/4 is related to cancer. The other feature of the Carcinogenic networks, regulation from Nanog to Nodal, was conserved among all of the Carcinogenic networks. Both Nanog and Nodal are important for normal early embryonic development. Nanog is a key factor for maintaining pluripotency in embryonic stem cells [33][34]. Nodal is related to the development of the left-right axial structure [35][36], and its signaling pathway is known to be important very early in development, for cell fate determination and many other developmental processes [36]. Although the Carcinogenic chemicals do not affect genetic structures, the regulatory mechanisms of these carcinogenic chemicals may be similar.

To compare the conserved gene relationships among chemicals with the same toxicity, we extracted the conserved gene regulations from the chemicals' networks. The numbers of conserved regulations were: 13 within Neurotoxic chemicals, 2 within Genotoxic chemicals, and 11 within Carcinogenic chemicals. Even though the average numbers of edges in the inferred models were similar among the three toxicity types (10.6 in Neurotoxic, 10.1 in Genotoxic, and 12.5 in Carcinogenic), the numbers of conserved regulations were different. From this feature, it is considered that a similar regulatory mechanism controlled the Neurotoxic chemicals' effects and the Carcinogenic chemicals' effects in ES cells, but the gene regulation by each Genotoxic chemical was independent of the toxicity type.

E. Similar mechanisms between chemicals

By utilizing the data mining method, we identified the chemicals with similar regulation. First, we constructed a transaction Table about the conserved regulation for each chemical, as shown in Table IV. Each row of data indicates the conserved regulation between genes, and each column indicates one chemical. In this transaction table, the value of 1 means that the corresponding regulation appeared with the chemical, whereas the value of 0 means that the regulation did not exist in the chemical's network.

In the affinity analysis, we set the thresholds as: Support > 0.5, Confidence > 0.5, and lift > 1. According to these restrictions, 2 rules were extracted. One is *BPA* => *DEAF*, and the other is *DEAF* => *PCB*. These results reflected the finding that the regulations in the *BPA* network were also conserved in the *DEAF* network. Furthermore, the regulations in the *DEAF* network were conserved in the *PCB* network. Although these three chemicals were categorized into different types of toxicities, they may share the same regulatory mechanisms to affect the ES cells.

2013, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

IV. CONCLUSION

We applied an improved SEM approach to reconstruct a gene regulatory model from gene expression data in human embryonic stem cells. Our results confirmed that SEM is a powerful approach to estimate the gene regulation caused by chemical toxicity. The shapes of the inferred network models for the various chemicals were different, but the inferred networks had a tendency to finally affect the same gene by their toxicity type. One of the neuron development-related genes, Tuj1, was arranged as the result of almost all of the Neurotoxic toxicity networks. Furthermore, Oct3/4 was important for both the Genotoxic and Carcinogenic networks. Since the Genotoxic chemicals are also carcinogenic, Oct3/4 is considered to be carcinogenic in ES cells. We detected some specific features for each toxicity type, and thus the inferred network among genes can be utilized for the estimation of a chemical's effects, 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 regulatory mechanisms by chemical toxicity.

TABLE IV. TRANSACTION TABLE OF CONSERVED REGULATIONS

edge info.		Neurotoxic chemicals					Genotoxic chemicals			Carcinogenic chemicals				Other chemicals		
parent	child	MeHg	2-Np	ACA	p-NA	PCB	BZP	DENA	DEAF	PB	TMX	DES	TCDD	THAL	BPA	PER
Oct34	Lmx1A	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
Oct34	Nestin	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0
Oct34	Pax6	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
Oct34	Tuj1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
GATA2	Oct34	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
GATA2	Lmx1A	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
GATA2	Nestin	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
GATA2	Pax6	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0
Lmx1A	GATA2	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0
Lmx1A	MAP2	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0
Lmx1A	Nanog	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
Lmx1A	Pax6	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
Lmx1A	Tuj1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
MAP2	Oct34	0	0	0	0	1	0	1	0	1	0	0	0	1	0	1
MAP2	GATA2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
MAP2	Lmx1A	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0
MAP2	Nanog	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0
MAP2	Nestin	1	1	0	1	0	0	0	1	0	0	0	1	0	1	0
MAP2	Pax6	0	0	0	1	1	0	0	1	0	0	0	0	1	1	0
MAP2	Tuj1	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0
Nanog	Oct34	1	0	0	0	1	0	0	1	1	0	1	0	0	0	0
Nanog	Lmx1A	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1
Nanog	MAP2	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
Nanog	Nestin	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
Nanog	Nodal	0	0	1	1	0	1	0	0	1	1	1	1	0	0	0
Nanog	Tuj1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
Nestin	Oct34	0	0	1	1	0	0	0	1	0	1	0	0	0	0	0
Nestin	GATA2	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
Nestin	Nanog	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
Nestin	Nodal	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0
Nodal	Oct34	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0
Nodal	MAP2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
Nodal	Nanog	1	1	0	0	1	0	1	1	0	0	0	0	0	1	1
Nodal	Nestin	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
Nodal	Pax6	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
Nodal	Tuj1	1	0	1	0	0	0	0	1	0	1	0	0	1	0	0
Pax6	Oct34	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0
Pax6	GATA2	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
Pax6	Lmx1A	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0
Pax6	MAP2	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1
Pax6	Nanog	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Pax6	Tuj1	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0
Tuj1	Oct34	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0
Tuj1	Nanog	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0
Tuj1	Nodal	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
Tuj1	Pax6	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1

a. The first column indicates the starting gene of one edge, and the second column indicates the end gene of the same edge.

b. The value of 1 means that the corresponding regulation appeared with the chemical, whereas the value of 0 means that the regulation did not exist in the chemical's network.

2013, © Copyright by authors, Published under agreement with IARIA - www.iaria.org