Prediction of Chemical Toxicity by Network-based SVM on ES-cell Validation System

Wataru Fujibuchi[1] **Sachiyo Aburatani**[1]
w.fujibuchi@aist.go.jp s.aburatani@aist.go.jp
Junko Yamane[2] **Satoshi Imanishi**[2]
yamane-j@m.u-tokyo.ac.jp imanishi@m.u-tokyo.ac.jp
Hiromi Akanuma[3] **Hideko Sone**[3]
akanuma.hiromi@nies.go.jp hsone@nies.go.jp
Seiichiroh Ohsako[2]
ohsako@m.u-tokyo.ac.jp

[1] Computational Biology Research Center, Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan
[2] Center for Disease Biology and Integrative Medicine, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8654, Japan
[3] Research Center for Environmental Risk, National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba 305-8506, Japan

Keywords: Bayesian network, kernel SVM, chemical toxicity, gene expression, stem cell

1 Introduction

Stem cell-based chemical informatics becomes increasingly important in biomedical fields, especially in drug discovery research. Here we describe our advanced bioinformatics technology to predict toxicities of chemicals including methylmercury (causes "Minamata" disease) and thalidomide (causes baby deformities) using the ES cell-based chemical validation system. In our analysis, we find that the Support Vector Machine (SVM)[1] prediction accuracy is increased by inferred gene networks when the weights of network edges are used as SVM features.

2 Method and Results

With RT-PCR gene expression data from ES cells exposed to 15 chemicals that are categorized into "neurotoxic", "tumor-genesis", and "others", we compared two SVM prediction methods: 1) gene expression data only 2) gene expression data + gene network edge weights. For RT-PCR analysis, we chose 9 (+1 internal standard) genes known as important to neuron development and performed RT-PCR experiments with 5 doses and 4 time points, yielding a total of 9x5x4=180 dimensions of data as SVM features. For network analysis, we start with the existing Bayesian network (BN) inference algorithm called TAO-Gen[2,3] and have improved it to the parallel (replica-exchange) version called RX-TAOgen, in order to obtain stable networks. A total of 9x9=81 gene-to-gene weights (β in Fig.1) on network edges are obtained and used as additional features for SVM. The whole scheme is shown in Fig. 1.

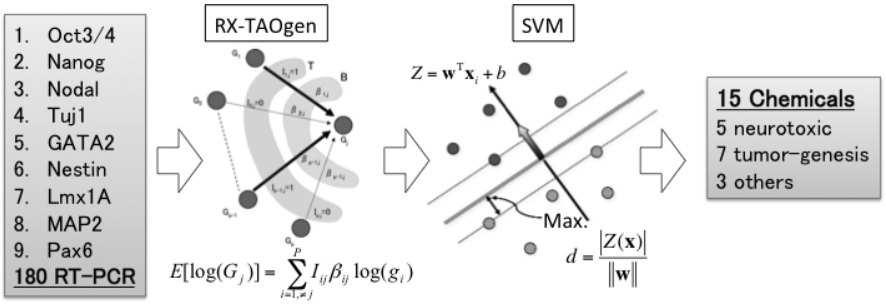

Fig.1 The whole scheme of toxic prediction using Bayesian networks as a part of input vectors of SVM.

2.1 Prediction Accuracy

For each of 15 chemicals, two RT-PCR experiments are performed, thus a total of 30 data are obtained. To evaluate the prediction accuracy, we performed a leave-two-out-prediction (LTOP) test, where two repeat data for each of 15 chemicals are used as test data at one cycle of prediction. In the SVM analysis, we tested three well-known kernels (linear, polynomial, and RBF) and our maximum entropy kernel[4] that is distance-based and the final kernel matrix is obtained by maximizing the kernel entropy. The accuracy is calculated by the ratio of the number of trues (true positives + true negatives) in all of the test data. The results are shown in Table 1.

Table 1: The number of data and accuracies for toxic chemical predictions.
(SVM: support vector machine, BN: Bayesian network)

Toxic category	Data (total of 30)	SVM	BN+SVM
Neurotoxic	10	90.0%	**93.3%**
Tumor-genesis	14	96.7%	**100.0%**
Others	6	–	–

2.2 Network Signatures

According to the results, the use of BN weights as additional features to SVM increases accuracies in both categories of toxic chemicals. The inferred BN structures are quite stable due to the high-tuned Gibbs-sampling method in RX-TAOgen; nevertheless, the obtained structures vary among chemicals. The network examples for 5 neurotoxic and 7 tumor-genesis chemicals are shown in Fig.2.

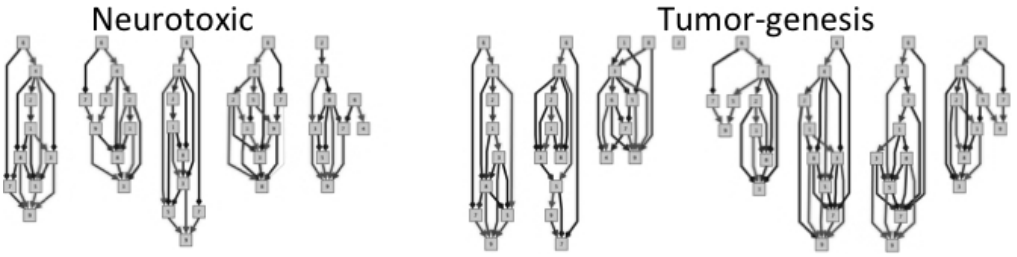

Fig.2 Examples of BN structures inferred from RT-PCR data for two chemical categories. Nine genes are indicated by rectangles and large weights are shown as edges.

3 Discussion

As the BN structures vary in Fig. 2 and it is difficult to recognize any distinct signatures in the same category by human eye, we performed a statistical comparison of the network weights of a pair of inferred networks. Interestingly, we observe that there are more correlations in weights among the same category than the different categories. For example, the average Pearson correlation coefficient of edge weights within 5 neurotoxic chemicals is 0.90, while that of 5 neurotoxic and 7 tumor-genesis chemicals is 0.86. Together, we conclude that the BN structures are weakly conserved in the toxic category, which may give additional information to SVM, resulted in the increase of accuracy.

References

[1] Vapnik, V., *The Nature of Statistical Learning Theory*, Springer, 1995.
[2] Yamanaka, T., Toyoshiba, H., Sone, H., Parham, F.M., and Portier C.J., The TAO-Gen algorithm for identifying gene interaction networks with application to SOS repair in E. coli., *Environ. Health. Perspect.*, 112:1614-21, 2004.
[3] Toyoshiba, H., *et al.*, Inference for Bayesian network via Gibbs sampling, Pre-print.
[4] Fujibuchi, W. and Kato, T., Classification of heterogeneous microarray data by maximum entropy kernel., BMC Bioinformatics, 8:267, 2007.