# Sparse modeling for sample-specific analysis

Heewon Park[1], Seiya Imoto[2] and Satoru Miyano[2]

[1]Faculty of Global and Science Studies, Yamaguchi University

[2]Institute of Medical Science, University of Tokyo

Over the last few dedicates, various statistical strategies have been proposed to understand the heterogeneous system of cancer. Although various methods have been developed to infer gene regulatory networks in cancer progression, the existing methods, such as L1-type regularization approaches, provide average modeling results for all samples, and thus we cannot reveal sample-specific characteristics of cancer based on the results from the existing methods.

We propose a novel statistical strategy for sample-specific analysis, called sample-specific stability selection (SS-stability selection), in line with the NetworkProfiler (Shimamura et al., 2011). The NetworkProfiler groups samples according to specific cancer characteristic by using the Gaussian kernel function, and performs modeling a target sample based on the grouped neighborhood around the target sample.

In order to effectively perform sample-specific analysis, we consider statistical modeling based on random lasso (Wang et al., 2011) and stability selection (Meinshausen and Bühlmann, 2010). By using the random forest method in random lasso, we can overcome the drawbacks of the existing L1-type regularization methods from multicollinearity, since only a small set of highly correlated variables may be considered as predictors in each bootstrap regression modeling. Furthermore, we consider a weighted bootstrap technique in random lasso, and thus our method effectively performs sample-specific analysis, because modeling for a target sample is based only on samples having similar cancer characteristic with a target sample. In short, we construct bootstrap datasets for a target sample having a specific cancer characteristic based on randomly selected predictors via the random lasso procedure, and then apply the NetworkProfiler to infer gene regulatory networks for the target sample. We finally perform feature selection based on the results of bootstrap regression modeling via modified stability selection procedure, and we show that our method can effectively control the false positive rate of feature selection in regression modeling.

## REFERENCES

[1] Meinshausen N. and Bühlmann P. (2010), Stability selection. *J. R. Stat. Soc. Ser. B,* 72, 417–473.

[2] Shimamura, T., Imoto, S., Shimada, Y., et al. (2011), A novel network profiling analysis reveals system changes in epithelial-mesenchymal transition. *PLoS ONE*, 6(6), e20804.

[3] Wang, S., Nam, B., Rosset, S., et al. (2011), Random lasso. *Ann. Appl. Stat.*, 5, 468-485.