# Combining Multiple Models in Reconstructing *In Silico* Regulatory Networks

Kevin Y. Yip[1], Roger P. Alexander[2,3], Koon-Kiu Yan[2] and Mark Gerstein[1,2,4]

*[1]Department of Computer Science; [2]Department of Molecular Biophysics and Biochemistry; [3]Department of Molecular, Cellular and Development Biology; [4]Program in Computational Biology and Bioinformatics, Yale University.*

We performed computational reconstruction of the *in silico* gene regulatory networks in the DREAM3 Challenges. Our task was to learn the networks from two types of data, namely gene expression profiles in deletion strains (the 'deletion data') and time series trajectories of gene expression after some initial perturbation (the 'perturbation data'). In the course of developing the prediction method, we observed that the two types of data contained different and complementary information about the underlying network. In particular, deletion data allow for the detection of direct regulatory activities with strong responses upon the deletion of the regulator while perturbation data provide richer information for the identification of weaker and more complex types of regulation.

We applied different techniques to learn the regulation from the two types of data. For deletion data, we learned a noise model to distinguish real signals from random fluctuations using an iterative method. The key idea is to first identify a conservative set of regulatory events that are unlikely to contain false positives. Wild-type expression levels and the width of the background Gaussian noise are then learned from the data outside this set of regulatory events. These learned values provide a probabilistic estimate of the chance that a given observed fluctuation is due solely to noise. The ones with very small probabilities then constitute a refined set of potential regulatory events. The whole process was repeated a number of times, which finally produced a probability for each gene A to be regulated by another gene B. Some ambiguous cases were resolved by checking the consistency between expression profiles in null mutants and heterozygous strains.

For perturbation data, we used differential equations to model the change of expression levels of a gene along the trajectories due to the regulation of other genes. We combined the predictions of various models (linear, sigmoidal and multiplicative). Model parameters were learned by using Newton's method with multiple starting points and the differential equations solved numerically by Runge-Kutta method. If a set of potential regulators result in a small squared difference from the observed expression values, they are likely the real regulators of the target gene. Due to the high computational cost, and to avoid overfitting, we started with models involving only one or two potential regulators. We then performed guided model fitting by using the obtained high-confidence set and the predictions from deletion data to construct more complex models.

The final predictions were obtained by combining the results from the two types of data. A comparison with the actual regulatory networks suggests that our approach is effective for networks with a range of different sizes.

To apply our method to real datasets, additional practical issues such as indirect regulation need to be confronted. Also, instead of performing model fitting in a purely unsupervised manner as described above, with the availability of some known examples of real regulatory networks, supervised or semi-supervised approaches could potentially lead to better performance.