Elucidating Transcriptional Regulatory Networks from heterogeneous gene expression compendia

Andrea Pinna[1], Nicola Soranzo[1], Vincenzo de Leo[1,2], Alberto de la Fuente[1]

[1]*CRS4 Bioinformatica, Loc. Piscina Manna, 09010 Pula (CA), Italy;* [2]*Linkalab, Complex Systems Computational Laboratory, 09100 Cagliari (CA), Italy*

The yearly organized DREAM challenges enable a fair comparison of methods for the inference of gene networks. Nowadays a constantly increasing amount of gene expression data is becoming available to researchers, so new techniques are needed for analyzing such huge and heterogeneous compendia. The goal of the DREAM5 Network Inference Challenge is to reverse engineer transcriptional regulatory networks from such datasets.

Challenge participants were given four microarray compendia (for three unspecified microorganisms and one *in silico* network), each one consisting of hundreds of microarray chips from different laboratories and with various combinations of genetic, drug and environmental perturbations. Knowing the experiment characteristics and the anonymized lists of transcription factors (TFs), participants are requested to rank all possible TF-target interactions.

In order to exploit the different types of information included in these mixed datasets, our team applied three distinct approaches and then combined them in a final prediction:
1) Perturbation-response analysis. If a TF is perturbed by knockout or over-expression in some experiment, we score each TF-target edge by the normalized deviation in the expression of the target gene after the TF perturbation with respect to the wild-type chip.
2) Partial correlation. Using the whole dataset except the time series, we calculated the full-order partial correlation for each (TF, target) pair.
3) Co-deviation analysis of aspecific perturbation data. We restricted this analysis to the chips with drug perturbations and 'non-TF single-gene perturbations'. For each TF, we calculated the two subsets of chips where the TF is 'high' and 'low', then scored each TF-target edge by the t-statistic of a two-sample t-test verifying whether the target expression mean is different between the two subsets.
We combined these scores by using the weighted mean of the edge ranks from the three different approaches. The weights for the average were obtained through studies on simulated data generated by our in-house simulator, which we adapted to perform simulations of the experiments considered in this challenge.

In the evaluation of the challenge by the organizers, our submission obtained an average overall score (10th out of 29), but with very different results on the real and synthetic datasets. In fact, our predictions were very good (2nd position) for the microorganisms, but not as much for the *in silico* network. In particular, our team was ranked 1st for the most difficult *S. cerevisiae* transcriptional network, the only eukaryotic one.

These results reveal that the combination of different inference techniques is indeed useful, especially with heterogeneous compendia, provided that these data are correctly subdivided. Moreover, even if gene expression simulations are more and more realistic, it clearly emerges that inference methods having good performances on synthetic datasets can not yet be expected to obtain the same results on real microarrays.