Gene Network inference by significance analysis on genotype/phenotype data

Tiziana Sanavia[1], Francesco Sambo[1], Angela Grassi[1], Barbara Di Camillo[1], Gianna Toffolo[1]

[1]*Department of Information Engineering, University of Padova, Italy.*

"DREAM5 SYSGEN A – In silico network challenge" investigates the use of genotyping and expression data for elucidating causal networks among genes. These data are provided for in-silico populations, where each gene exhibits a single DNA polymorphism either in the promoter region (cis-effect) or in the coding region (trans-effect). In the cis-effect case, two possible genetic variants, coded by 0 or 1, affect the gene expression at steady-state by a multiplicative factor of either 1 or 0.75.

Genetic polymorphisms can be interpreted as multifactorial perturbations that, combined with expression data, can be used to gain a global understanding of biological networks. With this purpose, we developed a method that relies on differential expression analysis of the data with respect to genetic variants. The method is based on two main steps:

1) Identification of the type of polymorphism: for each gene i, two groups of subjects are defined according to the two genetic variants (0 or 1) of i and significance analysis of microarrays (SAM) [1] is applied to detect the presence of a significant difference between the two groups for gene i itself. The rationale is, according to the model provided with the data, that if gene i is characterized by a cis-effect, its expression level depends on the genetic variant, whereas, if i is characterized by a trans-effect the genetic variant affects only the expression level of the genes regulated by i. Thus, a low p-value resulting from the test is highly indicative of the presence of a cis-effect. Moreover, the difference between the mean expressions in the two groups indicates which genetic variant affects the gene expression at steady-state by a multiplicative factor of 1 and which by a factor of 0.75. The expression values of these latter genes are divided by 0.75 before applying step 2, to remove the bias induced by the cis-effect.

2) Identification of causal regulatory effects: once the cis-effect has been identified and corrected as described above, for each possible regulating gene j, two groups of subjects are again defined according to the two genetic variants of j and SAM is applied to detect which genes are significantly differentially expressed. These genes are the candidate targets of the regulatory effect of gene j. The rationale is that, independently on the type of polymorphism, the effect of the genetic variant of a regulator is observable on its targets. Low p-values thus correspond to high confidence on regulatory effects. Results are ordered according to increasing p-value; in case of ties, predictions are ordered by correlation between the target genes and their regulator j.

Applied on the simulated data provided in the challenge, the method was proven highly reliable in inferring regulatory relations, as evidenced by the high mean AUROC: 0.68, 0.77, 0.84 for datasets of 100, 300 and 999 subjects, respectively.

[1] Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. PNAS 98: 5116-512.