Predicting causes and effects in regulatory networks

Marc Hulsman, Marcel J.T. Reinders
Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

**Problem description**

Connecting regulators with the genes that they regulate is difficult. While various large scale data sources are nowadays available (expression, literature, chip-chip, protein interactions, etc.), each of these is in itself an incomplete view of the regulatory network.

Our goal is to integrate them. We build a network that is able to predict how a perturbation in a gene S will affect other genes T.

**Approach**

Various (high-throughput) genome-scale data sources, containing information on different types of cellular networks, are used to determine the connectivity between source and target gene:

- Perturbation:
  - Protein-protein interactions
  - Gene expression data
  - Protein-DNA binding

- Causal interaction prediction
  - Protein-protein interaction
  - Transcription factors
  - Phosphorylation

Using this predictor, we want to be able to:

- predict gene expression effects (e.g., gene state changes)
- predict not affected genes
- predict unmeasured effects (Gi) in the network
- find genes which form a robust group of markers (Tj) for a certain perturbation (S)

**Method**

Network construction and data integration parameters

1) Network construction
- optimize network construction
- evolutionary strategy (CMA-ES)
- optimizing ROC-based score

2) Network simulation
- calculate probability of a gene to be affected by a defined knockout
- use message passing based Monte Carlo method

3) Network optimization
- optimize network construction
- evolutionary strategy (CMA-ES)
- optimizing ROC-based score

4) Validation on test set of knockouts
- using cross validation

**Results**

Using 10-fold cross-validation, the algorithm was validated on unseen gene perturbation microarray experiments.

**Discussion**

Building genome-wide, simulation models of cellular networks is difficult, as learning the activity of each of the millions of possible interactions easily leads to large numbers of false positives.

We simplify the problem, by learning a more general rule on how to integrate various data sources, using them to determine the regulatory activity of the interactions, thereby enabling predictions of causes/effects in regulatory networks.

We hope to transfer this knowledge to other species, enabling us to make maximally use of the available information in inferring regulatory networks.