

STATISTICAL INFERENCE OF BIOLOGICALLY-PLAUSIBLE DYNAMIC REGULATORY NETWORKS WITH CORE-LEAF TOPOLOGY

Michael O. Duff¹, Darya Chudova¹, Barbara Wold², Padhraic Smyth¹,
Eric Mjolsness¹

Abstract. Inferring regulatory networks from data is an important but challenging problem, and a number of formulations have been advanced, including Bayesian networks, Boolean switching networks, and linear networks. However, biological realism calls for an approach that incorporates a full combination of dynamics, nonlinearity, stochasticity, and known aspects of the underlying biological network topology. In addition, the sparsity and noisiness of current observed data make the use of statistical inference techniques essential in this context. Motivated by these factors, we describe a class of regulatory network models and associated novel statistical inference algorithms that learn both network structure and model parameters for networks with nonlinear, stochastic, regulatory dynamics. We present results from simulation experiments and relate them to real biological systems.

Specifically, we propose a *core-leaf* topology consisting of a densely-connected core of elements, together with leaf elements, each of which is connected to a small number of core elements (in this study, just one per leaf element). In this model, quantities represented by nodes in the network are governed by a system of stochastic difference equations of the form:

$$X_j^{(t+1)} = X_j^{(t)} + \alpha_j \mathcal{F}_{sigmoid} \left(\sum_i w_{ij} X_i^{(t)} + b_j \right) - \lambda_j X_j^{(t)} + \epsilon_j^{(t)}$$

where α_j , λ_j , and ϵ_j are node-dependent growth and decay parameters, and noise term, respectively, and where w_{ij} and b_j signify node connection strengths and bias. We use a statistical learning (Gibbs-sampling) approach to infer systematically both the topology of the core-leaf network and the network's dynamical system parameters. In each phase of Gibbs-sampling, the current network is subjected to a round of structural perturbations (each leaf becomes connected to a different core-parent or swaps roles with its core-parent); for each perturbed structure, associated dynamical system parameters are optimized against the data and produce a score (log-max-likelihood plus log-prior). We then sample from these scored, perturbed structures to produce a new network for the next phase of structural perturbations.

For a small (Repressilator-like) core-leaf network with three core nodes and three leaf nodes, in which there are 540 distinct candidate structures, we show successful identification of the true network using data generated at 25 time points. We investigate lower-bounds on the amount of data required for preserving identifiability, and alternatively, how identifiability can be enhanced by incorporating, in a principled way, alternative types of data (*e.g.*, data obtained from knockout experiments). We scale up our experiments to larger systems, present empirical results, and discuss links to biological and biologically-engineered systems with periodic behavior.

¹Institute for Genomics and Bioinformatics, University of California, Irvine.

²Division of Biology, California Institute of Technology.