Methods for gene regulatory network inference focus on network architecture identification but neglect model selection and simulation. We implement an extension to the dynGENIE3 algorithm that accounts for model uncertainty as an R package, providing users with an easy to use interface for model selection and gene expression profile simulation. Source code is available at https://github.com/tianyu-lu/dynUGENE with a detailed user guide. A webserver with interactive controls is available at https://tianyulu.shinyapps.io/dynUGENE/.

Introduction

Complex phenomena such as cell development and apoptosis emerge from coordinated dynamics of gene regulatory networks (GRN). Inferring network structure from data can be used for hypothesis generation, revealing mechanisms in cell development and disease (Huang et al., 2009), and modelling network evolution (Crombach and Hogeweg, 2008). Accurate dynamical models allow us to predict the effects of network perturbations on biological function, for example to push cells out of a disease state (Karlebach and Shamir, 2010), or to design synthetic GRNs given the desired dynamics of a network (Hiscock, 2019). The ideal model should be flexible enough to capture highly nonlinear interactions while not sacrificing model interpretability and computation time.

We present dynUGENE (dynamical Uncertainty-aware GEne NEwork inference), an R package that extends the functionality of dynGENIE3, a state-of-the-art method for GRN inference (Geurts et al., 2018). We build on dynGENIE3 because it satisfies all three of our model desiderata. Existing extensions include TIMEOR and BENIN which both incorporate heterogeneous data to improve network inference accuracy (Wonkap and Butler, 2020; Conard et al., 2020). Here, we take a different approach and instead account for uncertainty in dynGENIE3, allowing for stochastic gene expression simulations and parsimonious model selection. Our extension is available as an easy to use R package and also as an interactive web server.

Package Design

dynGENIE3 Background. dynGENIE3 poses GRN inference as a feature selection problem. It first trains random forests to predict the change in concentration of each species given the current concentrations of all species. Each interaction from species $x_i$ to species $x_j$ is associated with an importance score, calculated by the reduction in variance from using $x_i$ to predict the change in $x_j$. The importance score for an interaction, when normalized, is interpreted as the probability of that interaction to exist. For a detailed treatment, see the vignette and (Geurts et al., 2018).

Model Selection. The inferred network can be visualized as a $p \times p$ matrix where the entry $[x_i, x_j]$ is the importance score of $x_i$ for inferring $x_j$ (Fig. 1). However, real GRNs are often not fully connected and the presence of an interaction is binary (Mangan et al., 2016). To address this, dynUGENE includes a function for model selection based on visualizing the Pareto front (Mangan et al., 2016). However, we note that the model at the sharp drop in the Pareto front is not always the best model (Supplementary Fig. S1). We include an additional function on the web server where users can choose which interactions to mask. The masked networks can then be simulated, allowing for application-specific tuning of model complexity.

Model Simulation. The inferred networks and masked networks can be used to simulate gene expression profiles by numerically solving the system of ordinary differential equations learned by the random forests. In addition to deterministic simulations, we provide an option that accounts for the uncertainty in the random forests predictions for stochastic simulations. For stochastic simulations, instead of only taking the mean of a random forest’s predictions, we sample from the Gaussian $\mathcal{N}(\mu, \sigma^2)$ where $\mu$ is the mean and $\sigma^2$ is the variance of the random forest’s predictions.

Provided Datasets. The dynUGENE package provides four example time-series datasets: repressilator, stochastic repressilator, Hodgkin-Huxley, and stochastic Hodgkin-Huxley (Elowitz and Leibler, 2000; Hodgkin and Huxley, 1952). These datasets were generated from systems of ordinary or
stochastic differential equations. Details are provided in the vignette. The package also includes one steady state dataset, SynTReN300, taken from GRNdata (Bellot et al., 2020). Users can provide their own data as input following the for-...