Simulation of molecular regulatory networks with graphical models

Inma Tur¹, Alberto Roverato², Robert Castelo¹,*

1. Universitat Pompeu Fabra
2. Università di Bologna
*Contact author: robert.castelo@upf.edu

Keywords: Molecular regulatory network, Graphical model, Covariance matrix, Simulation

High-throughput genomics technologies in molecular biology produce high-dimensional data sets of continuous and discrete readouts of molecules within the cell. A sensible way to scratch at the underlying complex network of regulatory mechanisms using those data is to try to estimate the graph structure $G$ of a graphical model (Lauritzen, 1996). A fundamental step taken by many of the contributions to this problem is to test first the performance of the proposed algorithms on data simulated from a graphical model with a given graph $G$, before showing the merits of the approach on real biological data.

Here we introduce the functionality available in the R/Bioconductor package qpgraph (Tur et al., 2013) to simulate Gaussian graphical models, homogeneous mixed graphical models and data from them. The former produce multivariate normal observations which can be employed to test algorithms inferring networks from gene expression data, while the latter produce mixed discrete and continuous Gaussian observations, which can be employed to test algorithms inferring networks from genetical genomics data produced by genotyping DNA and profiling gene expression on the same biological samples.

A basic component to this functionality is the generation of a covariance matrix $\Sigma$ with: (1) a pattern of zeroes in its inverse $\Sigma^{-1}$ that matches a given undirected graph $G = (V, E)$ on $p = |V|$ vertices associated to $X_1, \ldots, X_p$ continuous Gaussian random variables; and (2) a given mean marginal correlation $\rho$ for those pairs of variables connected in $G$. This is achieved by applying a matrix completion algorithm (Hastie et al., 2009, pg. 634) on a $p \times p$ positive definite matrix drawn from a Wishart distribution whose expected value is determined by $\rho$ with $-1/(p-1) < \rho < 1$ (Odell and Feiveson, 1966). Building up on this feature, the package can interpret this matrix as a conditional one $\Sigma \equiv \Sigma(i)$, given a probability distribution on all joint discrete levels $i \in I$, and simulate conditional mean vectors $\mu(i)$ with given linear additive effects, which enable simulating homogeneous mixed graphical models. Using the mvtnorm package, conditional Gaussian observations are simulated accordingly. This functionality is also integrated with the one of the qtl package for generating genotype data from experimental crosses to enable the simulation of genetical genomics data under some of the genetic models available in qtl. Critical parts of the code are implemented in C language enabling the efficient simulation of graphical models involving hundreds of random variables.

The technical complexity behind all these features is hidden to the user by means of S4 classes and methods that facilitate the simulation of these data, as illustrated in the vignette included in the qpgraph package (Tur et al., 2013) and entitled “Simulating molecular regulatory networks using qpgraph”.

References


