Gene expression

simPATHy: a new method for simulating data from perturbed biological PATHways

Elisa Salviato\textsuperscript{1,}\textsuperscript{*}, Vera Djordjilovi\textsuperscript{c2}, Monica Chiogna\textsuperscript{2} and Chiara Romualdi\textsuperscript{1,}\textsuperscript{*}

\textsuperscript{1}Department of Biology and \textsuperscript{2}Department of Statistical Sciences, University of Padova, 35121 Padova, Italy

*To whom correspondence should be addressed.

Abstract

Summary: In the omic era, one of the main aims is to discover groups of functionally related genes that drive the difference between different conditions. To this end, a plethora of potentially useful multivariate statistical approaches has been proposed, but their evaluation is hindered by the absence of a gold standard. Here, we propose a method for simulating biological data – gene expression, RPKM/FPKM or protein abundances – from two conditions, namely, a reference condition and a perturbation of it. Our approach is built upon probabilistic graphical models and is thus especially suited for testing topological approaches.

Availability and Implementation: The simPATHy is an R package, it is open source and freely available on CRAN.

Contacts: elisa.salviato.2@studenti.unipd.it or chiara.romualdi@unipd.it

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Signaling pathways within a cell may be altered for various reasons. A mutation, a stress condition, a disease may change the expression of genes or the abundance of proteins. These changes, in turn, impact other elements of the cell leading to a general pathway dysregulation. The goal of most statistical methods for omic data analysis is the identification of such dysregulations. Indeed, several multivariate statistical approaches (called gene set analyses, hereafter GSA) have been proposed to identify groups of genes (rather than single genes) that are altered (in expression levels, protein abundances, methylations levels, etc.) in a specific condition. GSA use a priori biological knowledge to categorize genes into group of functionally related entities such as the Gene Ontology or biological pathways (Tarca \textit{et al.}, 2013). Among these methods those that showed promising results in terms of result interpretability are those called topological methods (for an extensive review see Mitrea \textit{et al.}, 2013). Topological pathway analysis exploits biological relationships featured in biological pathways in order to find coordinated expression changes inside a pathway coherent with a specific experimental condition.

Validation of GSA algorithms requires benchmark datasets for which dysregulations are known. Such datasets are generally not available. Some comparative studies used as benchmark real disease-specific datasets for diseases present in KEGG database (Kanehisa and Goto, 2000). The general idea behind this validation approach is that a sound gene set analysis on these disease-specific datasets should flag the corresponding disease pathways in KEGG as particularly significant. Although this could empirically surrogate the absence of benchmark datasets, there are some limitations: (i) pathway annotations are often incomplete and final results could be affected by different annotations; (ii) absence of a gold standard does not allow estimation of specificity, sensitivity, power and accuracy in different and defined experimental settings; (iii) the comparative analysis is limited to gene expression data while gene set analysis could be applied to a series of different omic data. Taking this into account, producing synthetic data mimicking real dysregulations to be used as benchmark datasets, appears to be the correct strategy to validate GSA tools.

Thus, here we present simPATHy (a method for simulating data from perturbed biological pathways), a new R package that produces synthetic datasets useful to validate GSA methods focused on
topological properties of biological networks. Given a graph structure representing a biological network, functions in simPATHy allow simulating data in two different conditions. The two conditions, assumed to share the same graphical structure, are referred to as reference and dysregulated condition. Dysregulation is defined as a change in the strength of the links between network nodes. Such changes can be interpreted as activation/deactivation of network connections.

As far as we know, currently there are no tools that simulate data according to this design. SynTREN (Van den Bulcke et al., 2006), one of the widely used network and data simulator, generates a biological network (subsampling nodes and edges from a real network following specific rules) and simultaneously simulates gene expression data derived from this network. Its main goal is regulatory network inference while in the context of gene set analysis the main issue is the identification of measurement differences between two or more groups.

2 Method

We assume to model the data of the same pathway in different experimental conditions as realizations of either an undirected Gaussian graphical model or a Gaussian Bayesian network sharing the same graphical structure G, see for instance Friedman et al. (2000) and Massa et al. (2010). Our method is applicable whenever biological data can be assumed to be Gaussian; these include, among others, log transformed microarray gene expression data, squared biological data can be assumed to be Gaussian; these include, among others, log transformed microarray gene expression data, squared root reads counts or RPKM/FPKM in next generation sequencing experiments, protein abundances. It modifies a covariance matrix of the reference condition in a manner that emulates signal dysregulation. Our proposed algorithm is able to: (i) obtain an estimate of a covariance matrix compatible with a given graph G starting from a sample covariance matrix, (ii) modify the strength of chosen pairwise correlations that emulate signal dysregulation, (iii) provide an adequate repair mechanism, based on the spectral decomposition of a correlation matrix, for indefinite matrices.

3 Example

3.1 Input

simPATHy as an input takes a graph and a covariance matrix associated to a reference condition. This covariance matrix might be a sample covariance matrix or any other estimate of the covariance matrix of variables in the reference condition. Another argument to the simPath function is a path in the graph that is to be dysregulated. By dysregulation we intend some modification of the vector of the pairwise correlation coefficients associated to edges of the chosen path (for more details see Supplementary material). To specify a path, a user can either provide directly a list of edges that constitute a desired path or choose a path in an interactive plot through the function getPathShiny. After specifying the path, a user can set the magnitude of the dysregulation. Each correlation coefficient pertaining to the path can be modified independently by providing the bounds (lower and upper) within which the actual multiplicative constant will be uniformly sampled. A value strictly between 0 and 1 will deactivate the signal, meanwhile, a value strictly greater than 1 will activate the signal. The user can also switch the direction of the relation (see Supplementary material).

3.2 Output

The output of simPATHy includes: (i) a dataset of observations coming from two conditions (reference and dysregulated) simulated from a multivariate normal distribution, (ii) covariance matrices of the two conditions and (iii) a summary of the parameters used for dysregulation. A quick look of the output can be obtained by the easyLookDys function.

3.2.1 Use of the simulated data

The main output is the simulated dataset of observations on two conditions that can be passed to GSA methods, such as Martini et al. (2013), Massa et al. (2010), Tarca et al. (2009) or Jacob et al. (2012), and used as benchmark. Some examples are provided in Supplementary material.

3.2.2 Visualization

simPATHy provides two functions for result visualization (focussing on the parameters of the two conditions represented by the graph and the covariance matrices), see Table 1. plotGraphNELD3 explores the graph in a 3D animation, allowing the user to move the nodes and zoom in parts of the graph. Colors of the edges represent the strength of the link (pairwise correlation or partial correlation) between two nodes. plotCorGraph allows the user to explore in more detail the correlation matrix (or the partial correlation matrix), while also showing the underlying graphical structure directly in the correlation matrix. Both of these plots can be explored dynamically by calling easyLookShiny, a Shiny app that gives a visual summary of the simPATHy output. For additional information, see documentation of the shiny package https://cran.r-project.org/web/packages/shiny/shiny.pdf.

Funding

This work was supported by Italian Association for Cancer Research (GI17185 to CR), by the University of Padova and by the Italian Ministry of Education, University and Research.

Conflict of Interest: none declared.

References