BIOINFORMATICS APPLICATIONS NOTE
Vol. 30 no. 4 2014, pages 571–573
doi:10.1093/bioinformatics/btt705

Gene expression
Advance Access publication December 3, 2013

Cascade: a R package to study, predict and simulate the diffusion of a signal through a temporal gene network

Nicolas Jung[1,2], Frédéric Bertrand[2,*], Seiamak Bahram[1], Laurent Vallat[1] and Myriam Maumy-Bertrand[2]

[1]INSERM UMR S_1109, Labex Transplantex, FMTS, Hôpitaux and Faculté de Médecine, Université de Strasbourg, 67085 Strasbourg Cedex and [2]IRMA, CNRS UMR 7501, Labex IRMIA, Université de Strasbourg, 67084 Strasbourg Cedex, France

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Summary: Temporal gene interactions, in response to environmental stress, form a complex system that can be efficiently described using gene regulatory networks. They allow highlighting the more influential genes and spotting some targets for biological intervention experiments. Despite that many reverse engineering tools have been designed, the Cascade package is an integrated solution adding several new and original key features such as the ability to predict changes in gene expressions after a biological perturbation in the network and graphical outputs that allow monitoring the spread of a signal through the network.

Availability and implementation: The R package Cascade is available online at http://www-math.u-strasbg.fr/genpred/spip.php?rubrique4.

Contact: fbertran@math.unistra.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 6, 2013; revised on October 4, 2013; accepted on November 28, 2013

1 INTRODUCTION

Since the emergence of high-throughput technologies that allow measuring simultaneously expression of thousands of genes, many tools have been developed to learn gene expression profiles and reverse engineer their underlying gene regulatory network (GRN) (Bar-Joseph *et al.*, 2012; Hecker *et al.*, 2009). These tools are either based on static coexpression methods or, if the biological phenomenon shows any temporality, time-dependent methods. Although the former relies on the assumption that coexpressed genes share some biological characteristics, the latter infers a directed network with temporal dependencies. In this last case, another important distinction should be made between exogenous stress (e.g. growth response) and endogenous phenomenon (e.g. cell cycle) (Yosef *et al.*, 2011; Zhu *et al.*, 2007). This leads to different network topologies: in exogenous stress, networks' topologies seem to have larger hubs and shorter paths through temporal-dependent transcriptional waves (Luscombe *et al.*, 2004). This results in a quick response to environmental modifications (Luscombe *et al.*, 2004). The Cascade package is designed to model such 'cascade networks' taking advantage of the assignment of genes to temporal clusters, which adds temporal causality in the network.

2 DETAILS ON THE PACKAGE FEATURES

This package has been designed to analyze temporal microarray datasets, allowing gene selection, temporal cluster assignment, reverse engineering the GRN using a penalized regression model and predicting the effect of biological intervention experiments. It also features a temporal synthetic cascade simulation tool. The biological interpretations are facilitated thanks to several graphical outputs. More insight about the statistical tools as well as benchmarks is provided in Vallat *et al.* (2013).

2.1 Gene selection and cluster assignment

Selecting the genes for reverse engineering is a crucial step. Besides selecting genes with high-differential expressions, the Cascade package allows enriching the selection with genes featuring specific temporal patterns. As pointed out by Hao and Baltimore (2009), several temporal gene expression waves, corresponding to specific cellular functions, can be individualized after stimulation of the cellular environment. In this pulsed biological response, some relevant genes may have low but systematic differential expressions. This selection step mostly relies on the Bioconductor R package limma (Smyth *et al.*, 2005).

Each gene must be then assigned to one of the time clusters. This can be automatically performed (according to the first time when the gene is differentially expressed). Alternatively, the time clusters can be user-provided.

2.2 Reverse engineering of the network

The reverse engineering algorithm is the Lasso-penalized estimation of a linear regression model described in Vallat et al. (2013). The Lasso penalty ensures sparsity, which is a well-known feature of most biological networks (Barabási, 2003). Furthermore, the temporal gene clusters are taken into account using a set of matrices F to describe how genes interact:

$$Y = \sum_{i=1}^{N} F_{m(X_i)m(Y)\omega_i} X_i + \eta \quad (1)$$

where Y is the regulated gene and the X_i are potential regulator genes, the ω_i determine the strength of the link between X_i and Y, $m(\cdot)$ is the function that maps a gene to its temporal cluster and η is

*To whom correspondence should be addressed.

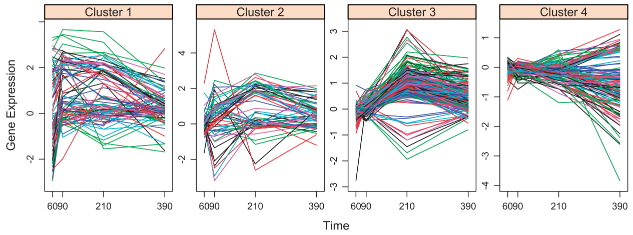

Fig. 1. Step 1: gene selection in GSE39411 and assignment to a time cluster

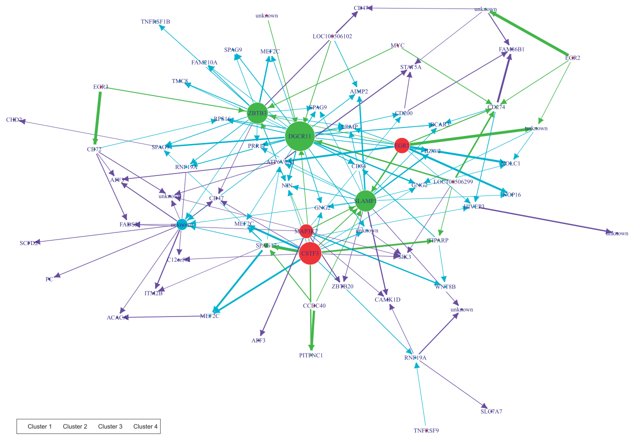

Fig. 2. Step 2: Reverse engineering of the network in GSE39411. Nodes represent genes and the arrows statistical links between the genes. Arrows' thickness depicts the intensity of the link

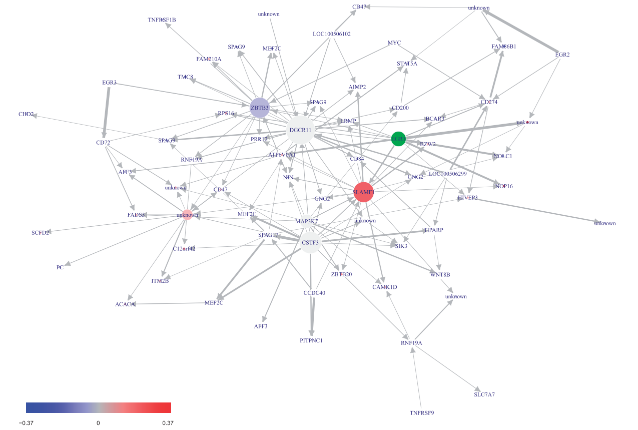

Fig. 3. Step 3: predicted perturbations in the network, at the second time point, after gene expression modulation at an early time in the temporal GRN of GSE39411. The green influential gene is supposed to be knocked down. Color scale legend from downregulated (blue) to upregulated (red) genes

a noise. Some further constraints are set to ensure a temporal causality, and we use the Lasso estimator to achieve some sparsity.

It is common knowledge that biological networks are scale-free (Barabási, 2003): the distribution of the outgoing edges in the networks follows a power law distribution. As a consequence, using a statistical test from Clauset *et al.* (2009), we derived a cutoff value for the coefficients ω. It was established, by a simulation study, that such a procedure greatly improves F-scores (Van Rijsbergen, 1979). A graphical output, Supplementary Material S1, shows the modification of the network topology when this cutoff varies. For a given cutoff, a graphical output, Supplementary Material S2, shows how the stimulated transcriptional response spreads through the network. If time clusters are heterogeneous, matrices F and ω values are iteratively estimated in a coordinate ascendant approach. On the contrary, if all the time clusters are homogeneous enough, the estimation of the matrices F may be achieved using all the genes in each of the clusters, instead of using only those pointed out by their ω values. This results in a non-iterative algorithm: matrices F and ω values are only estimated once.

2.3 Prediction

We can predict changes in gene expressions, using Equation (1), after a gene intervention experiment at the first time point, as validated, *in silico* and biologically in Vallat *et al.* (2013).

2.4 Simulation

The Cascade package provides two simulation tools. On the first hand, a random network can be simulated following the preferential attachment theory (Barabási, 2003) with some constraints to ensure that the result is a temporal cascade network. On the other hand, the model, Equation (1), can be used to simulate gene expressions from any given network.

3 EXAMPLES

Two package's vignettes detail the comprehensive analysis of two example datasets. A first dataset, extracted from GSE39411, is based on the transcriptional response of healthy lymphocytes B-cells after antigenic stimulation (Vallat *et al.*, 2007). The second dataset (E-MTAB-1475) has a different experimental design and is based on the transcriptional response of murine lymphocytes T-cells after an *in vitro* stimulation that sustains cellular differentiation (van den Ham *et al.*, 2013). In both cases, gene expressions measured at different time points after cell stimulation are used to select genes with specific temporal patterns or high differential expressions, which are then assigned to time clusters (Fig. 1 for GSE39411, and Supplementary Material S3 and S4 for E-MTAB-1475). The reverse engineering of the GRN highlights the most influential genes in the temporal cascade (Fig. 2 and Supplementary Material S3–S5). The impact in the GRN of a knockdown experiment of one influential gene is predicted (Fig. 3 and Supplementary Material S3–S6).

ACKNOWLEDGEMENT

Nicolas Jung, Frédéric Bertrand, Laurent Vallat and Myriam Maumy-Bertrand have contributed equally to this article.

Funding: Strasbourg High Throughput Next Generation Sequencing facility (GENOMAX), INSERM UMR_S 1109, CNRS UMR 7501, University of Strasbourg, INSERM

(ITMO cancer, Systems Biology, plan cancer 2009-2013); CNRS (PEPS-BMI interdisciplinarity 2013).

Conflict of Interest: none declared

REFERENCES

Barabási,A.L. (2003) Emergence of scaling in complex networks. In: Bornholdt,S. and Schuster,H.G. (eds) *Handbook of Graphs and Networks: From the Genome to the Internet*. Wiley-VCH, Weinheim, pp. 69–84.

Bar-Joseph,Z. *et al.* (2012) Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.*, **13**, 552–564.

Clauset,A. *et al.* (2009) Power-law distributions in empirical data. *SIAM Rev.*, **51**, 661–703.

Hao,S. and Baltimore,D. (2009) The stability of mRNA influences the temporal order of the induction of genes encoding inflammatory molecules. *Nat. Immunol.*, **10**, 281–288.

Hecker,M. *et al.* (2009) Gene regulatory network inference: data integration in dynamic models-a review. *Biosystems*, **96**, 86–103.

Luscombe,N.M. *et al.* (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.

Smyth,G.K. (2005) Limma: linear models for microarray data. In: Gentleman,R. *et al.* (eds.) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, pp. 397–420.

van den Ham,H.J. *et al.* (2013) Early divergence of Th1 and Th2 transcriptomes involves a small core response and sets of transiently expressed genes. *Eur. J. Immunol.*, **43**, 1074–1084.

Van Rijsbergen,C.J. (1979) *Information Retrieval*. 2nd edn. Butterworth, London.

Vallat,L. *et al.* (2007) Temporal genetic program following B-cell receptor cross-linking: altered balance between proliferation and death in healthy and malignant B cells. *Blood*, **109**, 3989–3997.

Vallat,L. *et al.* (2013) Reverse-engineering the genetic circuitry of a cancer cell with predicted intervention in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. USA*, **110**, 459–464.

Yosef,N. and Regev,A (2011) Impulse control: temporal dynamics in gene transcription. *Cell*, **144**, 886–896.

Zhu,X. *et al.* (2007) Getting connected: analysis and principles of biological networks. *Genes Dev.*, **21**, 1010–1024.