Bioinformatics, 33(5), 2017, 764–766
doi: 10.1093/bioinformatics/btw729
Advance Access Publication Date: 9 December 2016
Applications Note

Gene expression

LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering

Alicia T. Specht and Jun Li*

Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, USA

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on May 18, 2016; revised on October 21, 2016; editorial decision on November 11, 2016; accepted on November 15, 2016

Abstract

Summary: To construct gene co-expression networks based on single-cell RNA-Sequencing data, we present an algorithm called LEAP, which utilizes the estimated pseudotime of the cells to find gene co-expression that involves time delay.
Availability and Implementation: R package LEAP available on CRAN
Contact: jun.li@nd.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene co-expression networks (GCNs) use nodes to represent genes and edges to represent co-expression (simultaneous expression/silence, or simultaneously high/low expression) of genes, and they can be used to predict gene functions, among many other applications. Computational inference of GCNs are often based on a set of experiments from different tissues or different conditions, each measuring the expression of a large set of genes by high-throughput techniques like microarrays or RNA-sequencing (see e.g. Allen *et al.*, 2012; Specht and Li, 2015), the current *de facto* standard. A popular and most straightforward way of constructing a GCN is called the “correlation-based” approach, which connects gene pairs whose expressions in different biological samples are highly correlated, measured by Pearson’s correlation or other correlation coefficients.

Correlation-based GCNs constructed based on microarray or RNA-Sequencing data only capture simultaneous associations between pairs of genes. Biologically, if a gene enhances/inhibits another gene, then the latter gene will have delayed expression/silence (Munsky *et al.*, 2012). For such a pair, the co-expression of the two genes is strong if the delay in time is taken into account, but can be weak if only simultaneous association is considered. Unfortunately, such “time” information is not available in gene expression measured by microarrays or (bulk-based) RNA-Sequencing data, and thus correlation-based GCNs constructed using these data have limited ability to capture these regulatory relationships between genes, which are of equal, if not greater, interest than simultaneous expressions. A pioneering extension of the regular RNA-Sequencing technique, single-cell RNA-Sequencing (scRNA-Seq) is able to capture this time information in an indirect way. scRNA-Seq measures the gene expression profile of each individual cell, and hundreds to thousands of cells in a single run. These cells are at different time points of their cell cycles, and these time points can be estimated based on the idea that expression profiles are similar in cells at similar time points. These estimated time points are called “pseudotime,” and a few algorithms have been developed for the estimation (Campbell and Yau, 2015; Campbell *et al.*, 2015; Trapnell *et al.*, 2014; Reid and Wernisch, 2015).

2 Methods

We propose an algorithm called LEAP (Lag-based Expression Association for Pseudotime-series) for the computation of gene co-expression that takes into account the possible lags in time. LEAP sorts cells according to the estimated pseudotime (without considering branching), and then computes the maximum correlation of all possible time lags. This maximum correlation is used as the statistic to replace the traditional Pearson’s correlation coefficient for constructing the network, and the statistical significance of this statistic is measured by the false discovery rate calculated using permutations.

LEAP works by calculating the correlation of normalized mapped-read counts over varying lag-based windows. Given $X_{i,t}$ and $X_{j,t}$, the normalized and log-transformed number of reads mapped to genes i and j (e.g. $\log(\text{RPKM}+1)$ and $\log(\text{TPM}+1)$, where RPKM and TPM stand for reads per kilobase million and transcripts per million, respectively.), across experiments $t \in \{1, \ldots, T\}$ ordered by pseudotime, we examine windows of size s (we use $s = 2T/3$ for our real data example). For a given lag $l \in \{0, 1, \ldots, T-s\}$, we take the series $X_{i,0} = \{x_{i,1}, \ldots, x_{i,s}\}$ and $X_{j,l} = \{x_{j,l+1}, \ldots, x_{j,l+s}\}$, and find their Pearson's correlation $\rho_{ijl} = \frac{\text{cov}(X_{i,0}, X_{j,l})}{\text{std}(X_{i,0})\text{std}(X_{j,l})}$. We estimate this for all gene pairs across all $l \in L = \{0, 1, \ldots, T-s\}$, keeping the maximum absolute correlation (MAC) found, ρ^*_{ij}, where

$$\rho^*_{ij} = \max_{l \in L} |\rho_{ijl}|.$$

We use ρ^*_{ij} as the measurement of the strength of co-expression between gene i and gene j. By maximizing over all possible time lags, ρ^*_{ij} can often be larger than the regular measure of co-expression (Pearson's correlation coefficient without considering time lags). Another difference is that in general the gene pairs (i, j) and (j, i) do not have the same MAC, i.e. $\rho^*_{ij} \neq \rho^*_{ji}$. This is because ρ^*_{ij} measures the co-expression when gene j's expression is simultaneous or delayed compared to the expression of gene i. Thus, ρ^* is able to capture directional relationships. This directional relationship likely implies regulatory relationship: let $l^* = \text{argmax}_{l \in L}|\rho_{ijl}|$, then $\rho^*_{ij} > 0$ with $l^* \neq 0$ suggests gene i enhances gene j, $\rho^*_{ij} < 0$ with $l^* \neq 0$ suggests i inhibits gene j, and $l^* = 0$ suggests that gene i and gene j are both regulated by a third gene (Munsky *et al.*, 2012).

To measure the statistical significance of the ρ^* matrix we include a function to estimate the false discovery rate (FDR). We permute each gene i's normalized expression counts K times, creating $x_{i,t,0}, \ldots, x_{i,t,K}$. Then for each permutation k, we estimate ρ^*_{ijk}. For a cutoff C, the number of observed significant results is given by $N^{\text{observ}}_C = \sum_{i=1}^{n} I_{\rho^*_{ij} \geq C}$, where I is the indicator function, and the average number of significant results across K permutations is $N^{\text{perm}}_C = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n} I_{\rho^*_{ijk} \geq C}$. Finally, an estimate of the FDR is given by $\widehat{\text{FDR}} = N^{\text{perm}}_C / N^{\text{observ}}_C$.

Our implementation of LEAP using R is highly efficient. When calculating ρ^*_{ij}, we use a matrix computation with warm start to give ρ^*_{ij} for all gene pairs simultaneously. When doing permutations for estimating FDR, we randomly subsample genes at each permutation, which accelerates the computation dozens of folds with little loss of accuracy. For a dataset with 500 genes and 500 cells, LEAP takes about one minute to complete (with 100 permutations) on a regular laptop using a single core.

3 Results

To test LEAP's performance, we use a scRNA-Seq dataset that consists of 564 *Mus musculus* dendritic cells (Shalek *et al.*, 2014). We use $\log(x+1)$ transformed TPM (transcripts per million) values as the gene expression, and we refine the dataset to the most highly expressed genes using mean expression and relative interquartile range (IQR) cutoffs of 15 and 1.3, respectively, resulting in 557 genes. The pseudotime was estimated using Monocle (Trapnell *et al.*, 2014), which works by first mapping the gene expressions to low-dimensional space and then finding the longest path along a minimum spanning tree of the cell's locations. The resulting pseudotimes were kept for the 512 cells from the same state and used to sort the cells, compute ρ^*_{ij}, and estimate FDR using a set of thresholds $C \in (0, 1)$.

Table 1. Performance of LEAP on a real scRNA-Seq dataset

C	$N^{\text{total}}_{\text{simple}}$	$N^{\text{known}}_{\text{simple}}$	$N^{\text{total}}_{\text{LEAP}}$	$N^{\text{known}}_{\text{LEAP}}$	$N^{\text{known}}_{l^*>0}$	$N^{\text{known}}_{l^*>50}$	$\widehat{\text{FDR}}$
0.23	8494	1313	14 735	2394	911	526	0.002
0.22	9640	1508	20 405	3315	1661	1007	0.005
0.21	11 101	1751	28 843	4686	2818	1744	0.01
0.20	12 942	2022	41 778	6767	4691	2927	0.02
0.19	15 142	2367	59 556	9508	7204	4579	0.05
0.18	17 761	2804	82 424	12 989	10 446	6746	0.08

To check the ability of LEAP in detecting biologically true regulatory relationships, we use the *Mus musculus* network available through FunCoup, an online database that infers functional associations from publications (Schmitt *et al.*, 2013). For performance comparison, we also compute a regular Pearson-correlation-based network without considering time lags.

Table 1 shows for several cutoff values C, the number of identified associations and correctly identified known associations based on the FunCoup network by LEAP ($N^{\text{total}}_{\text{LEAP}}$ and $N^{\text{known}}_{\text{LEAP}}$), the number of non-zero time lags among these known associations ($N^{\text{known}}_{l^*>0}$), the number of time lags that are greater than 50 among known associations ($N^{\text{known}}_{l^*>50}$), and the estimated FDR ($\widehat{\text{FDR}}$). For comparison of performance, we also compute a regular Pearson-correlation-based network without considering time lags and give its number of identified and correctly identified known associations ($N^{\text{total}}_{\text{simple}}$ and $N^{\text{known}}_{\text{simple}}$). It is clear that LEAP discovers much more gene regulatory associations as it is able to take the time lag into account. For example, under FDR cutoff 0.05, LEAP discovers 9508, compared to 2367, known associations.

4 Conclusion

Regular correlation-based GCNs only describe simultaneous gene co-expressions. By using the time information that is virtually freely available in scRNA-Seq data, we developed a method LEAP that is able to capture associations that were hidden by the time lags. The asymmetric associations detected by LEAP more likely reflect regulatory relationships as they describe which gene follows another gene in expression. As an R package, LEAP is simple to use and computationally efficient. It also generates output compatible with popular analysis packages such as WGCNA (Langfelder and Horvath, 2008) to facilitate further inference based on the network.

Funding

This publication was made possible, in part, with support from the Indiana Clinical and Translational Sciences Institute, in part by UL1TR001108 from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. JL is also supported by the FRSP Initiation Grant from the University of Notre Dame.

Conflict of Interest: none declared.

References

Allen,J.D. *et al.* (2012) Comparing statistical methods for constructing large scale gene networks. *PloS One*, 7, e29348.

Campbell,K. *et al*. (2015) Laplacian eigenmaps and principal curves for high resolution pseudotemporal ordering of single-cell RNA-seq profiles. *bioRxiv*, 027219.

Campbell,K., and Yau,C. (2015) Bayesian Gaussian Process Latent Variable Models for pseudotime inference in single-cell RNA-seq data. *bioRxiv*, 026872.

Langfelder,P., and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Munsky,B. *et al*. (2012) Using gene expression noise to understand gene regulation. *Science*, **336**, 183.

Reid,J.E., and Wernisch,L. (2015) Pseudotime estimation: deconfounding single cell time series. *bioRxiv*, 019588.

Schmitt,T. *et al*. (2013) FunCoup 3.0: database of genome-wide functional coupling networks. *Nucleic Acids Res*., **42**, D821–D828.

Shalek,A.K. *et al*. (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, **510**, 363–369.

Specht,A.T., and Li,J. (2015) Estimation of gene co-expression from RNA-Seq count data. *Stat. Its Interface*, **8**, 507–515.

Trapnell,C. *et al*. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol*., **32**, 381–386.