Bioinformatics, 35(16), 2019, 2875–2876
doi: 10.1093/bioinformatics/btz007
Advance Access Publication Date: 8 January 2019
Applications Note

Systems biology

Transcriptogramer: an R/Bioconductor package for transcriptional analysis based on protein–protein interaction

Diego A. A. Morais[1], **Rita M. C. Almeida**[2] and **Rodrigo J. S. Dalmolin**[1,3,*]

[1]Bioinformatics Multidisciplinary Environment, Federal University of Rio Grande do Norte, Natal, RN, Brazil, [2]Institute of Physics and National Institute of Science and Technology: Complex Systems, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil and [3]Department of Biochemistry, Federal University of Rio Grande do Norte, Natal, RN, Brazil

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on May 8, 2018; revised on November 28, 2018; editorial decision on December 29, 2018; accepted on January 4, 2019

Abstract

Motivation: Several freely available tools perform analysis using algorithms developed to identify significant variation of gene expression individually. The transcriptogramer R package uses protein–protein interaction to perform differential expression of functionally associated genes. The software assesses expression profile of entire genetic systems and reveals which biological systems are significantly altered in case-control designed transcriptome experiments.

Results: R/Bioconductor transcriptogramer package projects expression values on an ordered gene list to perform topological analysis, differential expression and gene ontology enrichment analysis, independently of data platform or operating system.

Availability and implementation: http://bioconductor.org/packages/transcriptogramer.

Contact: rodrigo.dalmolin@imd.ufrn.br

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

1 Introduction

The available tools to assess gene expression, such as limma (Ritchie *et al.*, 2015) and DESeq2 (Love *et al.*, 2014), were designed to evaluate expression of individual genes. However, gene products are known to interact with each other to collectively perform biological functions and metabolic pathways (Li *et al.*, 2018). The transcriptogram (Rybarczyk-Filho *et al.*, 2011), a systems biology-based method to analyze transcriptomes, uses protein–protein interaction (PPI) to build an ordered gene list. Briefly, this method clusters genes by the probability of their products interact with each other, ordering them in one dimension. In other words, interacting genes are expected to get closer to each other on the ordered gene list. The ordered gene list is then used to take the average expression of functionally associated genes in a given window with settable radius. This strategy reduces gene expression data noise and improves data measure reproducibility (da Silva *et al.*, 2014). This Applications Note introduces a new R package for transcriptional analysis based on transcriptograms. The transcriptogramer R package is able to perform analysis on RNA-Seq and Microarray expression data by applying to transcriptograms the protocol adopted by limma package, known to perform well and fast under many circumstances (Conesa *et al.*, 2016).

2 Implementation

Transcriptogramer R package performs topological analysis, differential expression (DE) and gene ontology (GO) enrichment analysis. The workflow initially requires an ordered gene list and an edge list with gene connections. The package contains ordered gene lists for four species (*Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae* and *Rattus norvegicus*). Edge lists can be created from STRINGdb PPI. Using the initial inputs, it is possible to measure degree and clustering coefficient of individual genes as well as the same properties of functionally associated genes in a window segment of the ordered gene list. The software also calculates graph

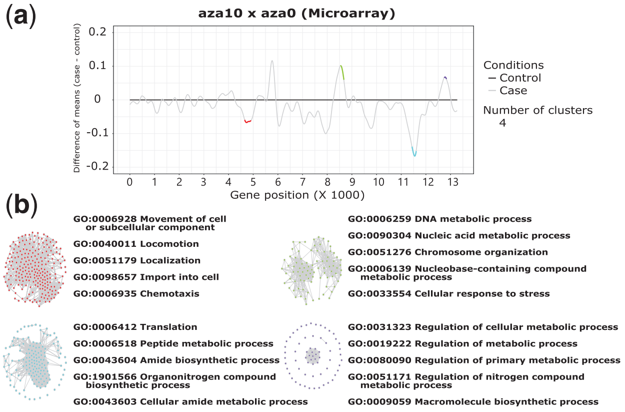

Fig. 1. (**a**) DE plot showing the difference of means between two conditions. Horizontal lines represent the average expression level of the conditions, black for the control samples and gray for case. Each highlighted area over the gray line represents one DEGG (False Discovery Rate < 0.02 for radius = 125). (**b**) DEGG PPI of significant genes used as windows centers and its top five terms (FDR < 0.05) taking into account GO hierarchy. Networks match the order of appearance in Fig. 1a from left to right, top down. A reproducible example can be found at https://github.com/arthurvinx/CaseStudy1

properties such as average assortativity and clustering coefficient as function of node connectivity. Transcriptograms are built in two steps that require expression data (Microarray or RNA-seq), a dictionary (mapping proteins to identifiers used as expression data rownames), and the initial inputs (the ordered gene list and the edge list). The first step assigns expression values (obtained from the expression data) to each respective gene in the ordered gene list. The second step assigns to each gene the average expression of neighbor genes in the ordered gene list. The logic is to define a sliding window centered on a given gene with a fixed radius (set by the user). The sliding window considers periodic boundary conditions to deal with genes near the ordered gene list borders. The goal here is to measure the average expression of functionally associated genes, represented by neighbor genes in the ordered gene list.

DE analysis evaluates expression variations of functionally associated genes in case-control experiments. As a result, the package provides a data.frame and plots a figure showing differentially expressed gene groups (DEGG) (Fig. 1a). PPI among inferred DEGG can be displayed as graphs (Fig. 1b) and it is also possible to perform GO enrichment analysis to identify the GO terms significantly associated with each DEGG (Fig. 1b). The complete transcriptogramer R package pipeline is shown in Supplementary Figure S1 and the main dependencies used by the package on the analysis are shown in Supplementary Table S1. The transcriptogramer package was implemented in R and is freely available at Bioconductor open source software for bioinformatics. The package is cross-platform, runs on Windows, Linux and Mac OS X. transcriptogramer is multi-thread compatible, allowing users to easily define the number of cores to be used on some methods, reducing runtime.

3 Application examples

The transcriptogramer R package was applied to Microarray and RNA-Seq datasets of human colorectal adenocarcinoma cells HT-29 treated with two different concentrations of 5-aza-deoxy-cytidine (NCBI-BioProject, PRJNA177602) (Xu *et al.*, 2013). The RNA-Seq reads were processed following a count-based protocol (Anders *et al.*, 2013) and log2-counts-per-million values were obtained using limma voom() function (the complete list of tools used in the protocol is shown in Supplementary Table S2 and the pipeline is shown in Supplementary Fig. S2). After processing, filtering, and normalization (a multidimensional scaling plot of the RNA-Seq samples is shown in Supplementary Fig. S3), data were analyzed using transcriptogramer package pipeline as well as another pipeline combining functions from limma and topGO (Alexa *et al.*, 2006) packages. Samples (9 at total) were classified into three groups (control, 5-aza-deoxy-cytidine 5 and 10 μM). Two case-control DE and functional enrichment were performed on Microarray and RNA-Seq data for each pipeline. All methodologies were able to identify similar number of differentially expressed genes. However, transcriptogramer R package was able to identify nearly ten times more GO terms when comparing with classic DE analyses followed by topGO enrichment in both Microarray and RNA-Seq data. A detailed comparison can be found in Supplementary Tables S3–S8.

4 Discussion

The transcriptogramer provides a systems biology-based pipeline for RNA-Seq and Microarray data transcriptional analysis, designed to identify DE of functionally associated genes. transcriptogramer advantage over existing functional analysis pipelines involves using canonical PPI data to select non-predefined gene sets, which enable to identify gene clusters responsible for a given GO term enrichment. Functionally associated genes have inherent hierarchical properties, thus, users can increase the window radius to merge gene sets and explore the enrichment results of bigger systems. The transcriptogramer R/Bioconductor package is an easy to install well-documented tool, developed to help bioinformaticians and biologists to convert their gene expression data into biological insights.

Acknowledgements

The authors would like to thank the NPAD/UFRN.

Funding

This work was supported by the CNPq [grant 444856/2014–5] and the CAPES [grant 3381/2013].

Conflict of Interest: none declared.

References

Alexa,A. *et al.* (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.

Anders,S. *et al.* (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.*, **8**, 1765–1786.

Conesa,A. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*

da Silva,S.R.M. *et al.* (2014) Reproducibility enhancement and differential expression of non-predefined functional gene sets in human genome. *BMC Genomics*.

Li,J. *et al.* (2018) Application of weighted gene co-expression network analysis for data from paired design. *Sci. Rep.*

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*

Rybarczyk-Filho,J.L. *et al.* (2011) Towards a genome-wide transcriptogram: the *Saccharomyces cerevisiae* case. *Nucleic Acids Res.*, **39**, 3005–3016.

Xu,X. *et al.* (2013) Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets. *BMC Bioinformatics*.