BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 10(4)

dna: An R package for differential network analysis

Ryan Gill[1], Somnath Datta[2] & Susmita Datta[2]*

[1]Department of Mathematics, University of Louisville, Louisville, KY 40292 USA; [2]Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40292 USA; Susmita Datta – Email: susmita.datta@louisville.edu; *Corresponding author

Received March 31, 2014; Accepted April 02, 2014; Published April 23, 2014

Abstract:
Differential network analysis provides a framework for examining if there is sufficient statistical evidence to conclude that the structure of a network differs under two experimental conditions or if the structures of two networks are different. The R package **dna** provides tools and procedures for differential network analysis of genomic data. The focus of this package is on gene-gene networks, but the methods are easily adaptable for more general biological processes. This package includes preprocessing tools for simultaneously preparing a pair of networks for analysis, procedures for computing connectivity scores between pairs of genes based on many available statistical techniques, and tools for handling modules of genes based on these scores. Also, procedures are provided for performing permutation tests based on these scores to determine if the connectivity of a gene differs between the two networks, to determine if the connectivity of a particular set of important genes differs between the two networks, and to determine if the overall module structure differs between the two networks. Several built-in options are available for the types of scores and distances used in the testing procedures, and additionally, the procedures provide flexible methods that allow the user to define custom scores and distances.

Availability: dna is freely available at The Comprehensive R Archive Network, http://CRAN.R-project.org/package=dna

Background:
In many genomic studies, it is important to examine the interactions/associations between genes and how these interactions change under different experimental conditions or differ between two populations. Various permutation-based statistical tests were presented in **[1]** for determining whether a gene or set of genes behaves differently in two networks. Each type of test is based on connectivity scores which measure the strength of the associations for each pair of genes in a network and on an associated distance function. In this paper, we describe the **dna** package which provides a versatile implementation of these tests with several options for connectivity scores and distance functions. The package is based on the open source R **[2]** software environment and freely available at **[3].** The built-in source code for performing the tests is primarily written in C to improve computational speed, but optional arguments are provided allowing the user to supply R functions for computing customized scores and distances. The function test.individual.genes identifies genes for which the connectivity scores differ significantly between the two networks, test.class.genes determines if there are significance differences in the scores for a subset of genes of interest while adjusting for all available genes, and test. modular.structure uses the scores to identify modules of genes that are connected and tests if the structures of the networks differ significantly using a statistic involving intersections and unions of the modules. Mathematical descriptions of the statistical tests are given in the vignette for the package.

The default connectivity score for the tests is based on partial least squares (PLS) regression using the algorithm in **[4].** There are other built-in options for connectivity scores including principal components regression, ridge regression, and the correlation coefficient. As mentioned before, users can define their own function for computing the connectivity score. A detailed example of a user-defined implementation of the LASSO based on the **lars** package **[5]** is provided in the vignette.

 © 2014 Biomedical Informatics

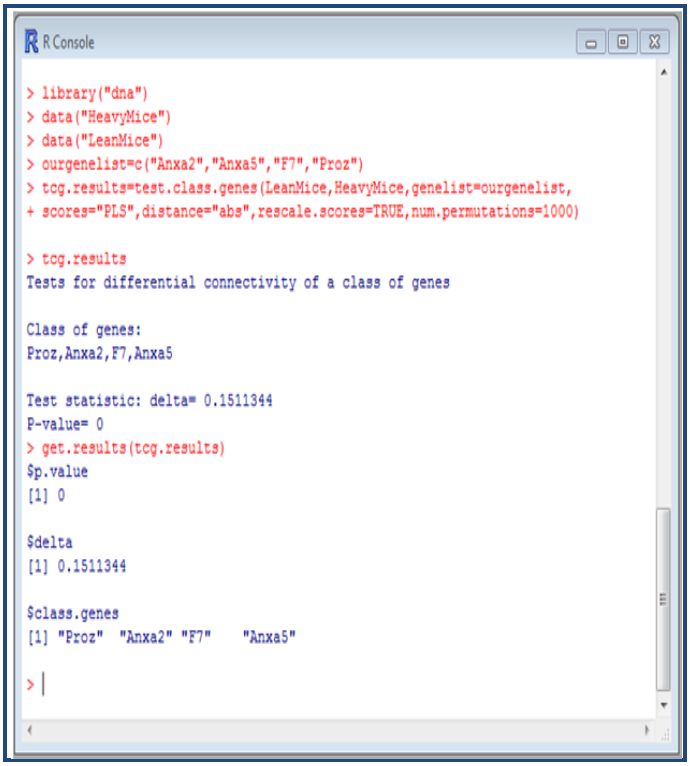

Figure 1: Usage and output from the **dna** package to perform a test for differential connectivity of the four genes Anxa2, Anxa5, F7, and Proz in a data set which includes genomic expression values for liver tissue from two groups of mice.

By default, the distance functions for test. individual.genes and test.class.genesis the absolute value of the difference between two scores; a built-in squared distance function is also available. Also, user-defined distance functions can be used, and an example of a piecewise distance function is provided in the vignette. For test.modular.structure, the user can select the minimum module size and the threshold connectivity parameter. A real data example based on mouse weight data previously analyzed in **[1], [6],** and **[7]** is included in the vignette. A screenshot of the test for differential connectivity of a class of four genes is shown in **Figure 1.**

Software input:
For each network, the user supplies a matrix of expression values in which the rows represent experimental units and the columns represent genes. If column names are provided, the package automatically selects only the genes common to both networks and aligns the corresponding genes in the custom S4 object used for differential network analysis. The user also has many options for preprocessing the data and calibrating the connectivity scores.

Software output:
The functions which perform the significance tests return objects, and there are summary and get. results methods for each type of object. For the test for individual genes, summary prints the number of genes which are significant at various levels and get.results returns a list of significant genes, their test statistic values, and their p-values. For the test for a class of genes, summary prints the test statistic value and p-value, and get.results returns a list of these values which can be accessed with R code. For the test for modular structure, summary prints the number of genes in the modules for each network, and get.results returns a list including the composition of each network, the test statistic, and p-value for the test. The functions for computing connectivity scores can also be accessed directly to create a score matrix. Additionally, there is a function network.modules which determines the modular structure of a network based on the connectivity scores. This function includes the option of displaying a graph of the network using the tkplot function from the **igraph** package **[8].**

Future development:
Several additional options will likely be made available in the **dna** package in the near future. Currently, regression models are used to compute connectivity scores based on modeling each gene's expression values based only on the expression values of the other genes; future versions will allow users to include other covariates to model each gene. In addition, the number of built-in methods for computing connectivity scores will increase. Finally, there are plans to develop specific objects for differential network analysis of different types of genetic data, such as objects specifically for next-generation sequencing data with discrete read counts.

Acknowledgement:
This work was partially funded by NIH/NCI Grant CA170091-01A1 to Susmita Datta.

References:
[1] Gill *et al. BMC Bioinformatics*. 2010 **11**: 95 [PMID: 20170493]
[2] http://www.R-project.org
[3] http://CRAN.R-project.org/package=dna
[4] Pihur V *et al. Bioinformatics* 2008 **24:** 561 [PMID: 18204062]
[5] http://CRAN.R-project.org/package=lars
[6] Fuller TF *et al. Mamm Genome*. 2007 **18**: 463 [PMID: 17668265]
[7] Ghazalpour A *et al. PLoS Genet*. 2006 **2**: e130 [PMID: 16934000]
[8] http://cran.r-project.org/web /packages /igraph/citation.html

Edited by P Kangueane
Citation: **Gill *et al.*** Bioinformation 10(4): 233-234 (2014)
License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited