TopoFunc: a machine learning method to identify functional modules in gene co-expression networks and complement Gene Ontology annotations

1 Introduction

TopoFunc searches for genes co-expressed with an initial set of functionally related genes, *e.g.* a Gene Ontology Biological Process (GO-BP). We first performed LASSO (least absolute shrinkage and selection operator) to select the topological descriptors of gene co-expression modules that discriminated modules made of functionally related genes (functional modules, FMs) and modules made of randomly selected genes (random modules, RMs). Using the selected topological descriptors, we performed Linear Discriminant Analysis to construct a topological score (ScoreTopo) that predicted the type of a module, random-like or functional-like. We combined the topological score and a functional similarity score (ScoreFunc) inspired by the work of Wang and co-workers (Wang, J.Z. *et al.* (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23, 1274-81) in a fitness function. Starting from a given GO-BP, we used a genetic algorithm to find genes whose co-expression with the largest clique of the GO-BP suggested that they may be functionally related. The TopoFunc manuscript provides further details about the algorithm.

2 Data files

2.1 graphALLdata.RData

An RData file that contains the co-expression data of 20,959 murine genes found in COXPRESdb Mmu.c3-1. It consists of 20,959 nodes and 10,901,364 edges with a mutual rank (MR) ≤ 1,200.

2.2 TrueGenesALLdata.RData

An RData file that contains the 20,959 GeneIDs.

2.3 MatriceGeneSim.RData

An RData file that contains the semantic similarity scores between the 20,959 genes. It is calculated from the Gene Ontology terms associated to the genes.

2.4 ResultsLDA.RData

An RData file that contains the results of the Linear Discriminant Analysis. These data are used to predict the type of a network, random-like or functional-like.

2.5 matrixGenesGOavant.RData

An RData file that records which of the 20,959 genes are part of the 193 GO-BPs studied in the Application section of the manuscript. It is composed of 20,959 rows (genes) and 193 columns (GO-BPs); each cell contains '1' or '0' depending on whether the gene is in the GO-BP or not.

3 Functions

3.1 ConstructionModule(gene_ids, threshold, path)

The co-expression data from COXPRESdb consist of 20,959 text files. Each file contains the MR values between 1 specific gene and the 20,958 other genes in COXPRESdb. The file's name is the GeneID of the specific gene. The file includes 3 columns and 20,958 rows.

- The first column contains the 20,958 GeneIDs.
- The second column contains the MR values between the specific gene and the 20,958 others.
- The third column contains the values of the Pearson correlation coefficient (PCC) between the specific gene and the 20,958 others.

The function extracts the subnetwork corresponding to a subset of genes and a MR threshold from COXPRESdb.

3.1.1 Argument

- gene_ids: a vector that contains the GeneIDs of the genes used to construct the subnetwork;
- threshold: the threshold value (µ in the manuscript) below which a MR is considered to denote co-expression. The default value is 1,200 (see the Supplementary File for a discussion about the default µ value);
- path: the path to the working directory containing the 20,959 files of the COXPRESdb.

3.1.2 Value

- RelationMatrix: a 3-column matrix that describes the edges of the subnetwork. Each edge is characterized by 2 nodes (GeneIDs) and 1 MR value. The 2 first columns contain the GeneIDs and the third column contains the MR value;
- AverageMutualRank: a 2-column matrix that contains the GeneIDs of the genes in the subnetwork and the corresponding average MR;
- Graph: an igraph graph.

3.2 ParametersPertinents(graph, pertinent=TRUE, Round=6)

This function calculates 12 topological descriptors of a network.

3.2.1 Argument

- graph: the 'graph' output variable from the ConstructionModule function;
- pertinent: a logical value, defaulting to TRUE, indicating whether the function calculates the relevant descriptors only or all descriptors;
- Round: an integer indicating the number of decimal places to be used in the calculation.

3.2.2 Value

Twelve descriptors calculated if pertinent=FALSE, and only the six relevant descriptors otherwise (see manuscript for the 12 descriptors).

3.3 LDAnormalization(scoretopo)

As mentioned in section 3.1.1 of the manuscript, ScoreTopo ranged from -2.535 to +6.466 on the learning data set. We introduced $\delta_0 = 0.1110948$ and $\delta_1 = 0.281608$ to normalize ScoreTopo by rescaling its value using the learning dataset so that the RM with the lowest ScoreTopo was set to 0 and the FM with highest ScoreTopo was set to 1.

3.3.1 Argument

- scoretopo: the value of ScoreTopo.

3.3.2 Value

The ScoreTopo value after normalization. It is the norm_scoretopo variable used in the fitness_function below.

3.4 fitness_function(norm_scoretopo, scorefunc, m1m0)

The fitness function takes as input 3 values characterizing a potential solution produced by the genetic algorithm and returns a numerical value measuring its ‘fitness’.

3.4.1 Argument

- norm_scoretopo: ScoreTopo value after normalization by the LDAnormalization function;
- scorefunc: ScoreFunc value;
- m1m0: the number of genes that belong to the module to evaluate, M1, and to the initial module, M0, divided by the number of genes in M0; card(M1∩M0)/card(M0)

3.4.2 Value

a numerical value that measures the ‘fitness’ of a solution.

3.5 plotTopoFunc(resAG)

This function gives a plot of the best values found during the iterations of TopoFunc.

3.5.1 Argument

- resAG: TopoFunc results.

3.6 TopoFunc (TrueGenesGO, TrueGenesALL, graphALL, ReduceSpace=FALSE, FunctionReducer=FALSE,
OptimiseInitialisation=FALSE, Ngene=100, aleatoire=FALSE,
TpopA = 100, TpopB = 300, TpopC = 100, Pmax=500,
FunctionParam=ParametersPertinents, FunctionLDA=ResultsLDA, distGO=MatriceGeneSim, LDAnorma=LDAnormalization, Criteria=criterion, tourn=2, pm=0.8, pc=0.5, AgeMax=10, plocImmig=1, conv=100
)

3.6.1 Argument

- TrueGenesGO: GO genes ID;
- TrueGenesALL: All genes ID;
- graphALL: the total network representing the relationship between all genes in database;
- ReduceSpace: a logical defaulting to FALSE determining whether or not we reduce the gene set. If the set of genes is reduced, the new set is built from the neighbors of GO's genes. In addition to the GO genes, the new set contains their neighbors having important links with the GO genes. (i.e larger than the total average of the number of links between GO genes and its neighbors);
- OptimiseInitialisation: a logical defaulting to FALSE determining whether or not we choose the first population from the reduced data set only;
- Ngene: the maximum number of iterations to run before the algorithm search is halted;
- aleatoire: a logical value specifying if the first generation-population is composed of *Tpop* random individuals (TRUE) or not (FALSE, default);
- *$Tpop_A$*: sub-population size of individuals identical to M_0. By default, 300;
- *$Tpop_B$*: sub-population size of individuals composed of the largest clique of M_0 and 80% of the genes of M_0 in the complement of the largest clique. By default, 100;
- *$Tpop_C$*: sub-population size of individuals composed of all genes in M_0 and (500 - card(M_0)) random genes (not in M_0). By default, 100;
- Pmax: the maximum size of genes in solution. By default, 500;
- FunctionParam: An R function performing selection, i.e. a function which calculate the topological parameters of network. By default, **ParametersPertinents**;
- FunctionLDA: the output of LDA function (lda(x, ...)) which seeks a linear combination of the relevant descriptors that optimally discriminated FMs and RMs. By default, **ResultsLDA**;
- distGO: a matrix containing the values of semantic similarity between each pair of genes;
- LDAnorma: An R function which normalizes ScoreTopo by rescaling its value using the learning dataset. By default, **LDAnormalization**;
- Criteria: the fitness function that returns a numerical value measuring the 'fitness' of a module with respect to the original module;

- selection: An R function performing selection, i.e., a function which generates a new population of individuals from the current population probabilistically according to individual fitness. The available functions are *Tournament* and *Rank*;
- tourn: number of individuals used in the tournament selection;
- pm: the probability of mutation in a parent chromosome. By default, 0.8;
- pc: the probability of crossover between pairs of chromosomes. By default, 0.5;
- AgeMax: maximum age of individual. By default, 10. After reaching a pre-defined 'AgeMax' (maximum age), the individual was removed from the population and replaced by a new individual composed of the elitism genes plus other genes from the rest of data;
- plocImmig: number of genes to add to the new individual above. By default, 1;
- conv: a number representing the maximum iteration after which the algorithm converge, i.e. we assumed that the algorithm had converged if the best solution was identical for 'conv' generations.

3.6.2 Value

- pop1: list of average *fitness* values of the population in each iteration;
- pop2: list of average *ScoreTopo* values of the population in each iteration; • pop4: list of average *Size* values

  of the population in each iteration;
- pop5: list of average *ScoreFunc* values of the population in each iteration;
- pop6: list of average *PercentageSize* values of the population in each iteration, where *PercentageSize* is the last term of criterion;
- genesCliqGO: genes of the largest clique of the initial module;
- Pmax: the maximum size of genes using in TopoFunc;
- iteration: maximum iteration used for TopoFunc to converge;
- bestfitness: the best fitness at each iteration;
- FMprime: the solution giving the best fitness at the final iteration;
- fitnessSummary: a matrix of the values of the criterion terms for the initial module and the final module.

4 Application

An application of the TopoFunc script is provided in the TopoFunc(GO_0006413).r file.

This script exemplifies the use of TopoFunc on GO:0006413 ('translational initiation'), which contains 52 genes whose GeneIDs are in column 27 of the matrix 'matrixGenesGOavant'.