

1. Description of the topological scoring algorithms

There are four algorithms available:

- (1) Shortest paths among nodes from user's list(default)
- (2) Transcription activation paths from drug targets to nodes in user's list
- (3) Shortest paths from all nodes to proteins in user's list
- (4) Transcriptional activation paths all nodes to proteins in user's list

The first (1) algorithm starts from each node from the input set and connects it through shortest paths to the rest of the nodes. For each node pair (one starting node from input set and a single node from the shortest path) the set of “reachable” nodes (via that node through shortest paths) will be identified and enrichment analysis with proteins from the input set will be performed. All nodes which become part of shortest paths will be assigned p values which will reflect the statistical significance of shortest path nodes providing connectivity among the proteins from the input set. The third (3) algorithm is similar to (1), the difference being that it starts from “all” nodes in the network (not just the input set) and connects them through shortest paths to the rest of the nodes and calculates enrichment in a similar way. In this case, we evaluate significance of shortest path nodes in providing connectivity from all nodes to the set of nodes from the input set. In (2) and (4) we build shortest path to transcription factor targets. In (2) we start from all nodes which are known drug targets and build shortest paths to all transcription factor targets. Only shortest paths which do not have an overall inhibition effect are considered. If a shortest paths contains multiple activation and inhibition interactions we use the following simple rule: the number of inhibition interactions should be zero or any other even number. In (4) we start from all nodes (not just the set of drug targets) when building shortest paths to transcription factor targets. (2) and (4) is different from the other two algorithm that we limit our shortest paths: (i) must end at a transcription factor target and (ii) must be not an “inhibition” path.

Overview:

(1)

input set (submitted) --> shortest path node --> enrichment (with input nodes) of the “reachable” set of nodes (p-in)

input set (submitted) <-- shortest path node ← enrichment (with input nodes) of the “reachable” set of nodes (p-out)

(3)

all nodes --> shortest path node --> enrichment (with input nodes) of the “reachable set” of nodes (p-in)

all nodes <-- shortest path node ← enrichment (with input nodes) of the “reachable set” of nodes (p-out)

For a detailed description of (1) see:

Z. Dezso, Y. Nikolsky, J. Miller, D. Cherba, C. Webb and A. Bugrim: Identifying disease-specific genes based on their topological significance in protein networks. BMC Systems Biology 2009, 3:36

2. Input files:

All 4 algorithms need a list of input genes (entrez genes id or affy id's) submitted as a column in a plain text file. A background list can also be submitted (affy or entrez gene id's). The background list will be used for the p value calculation. In case no background list was submitted the algorithm will use the set of nodes from the global protein interaction network as a background for the p value calculation. The p value submitted must be between 0 and 1 (if not submitted or is not a number between 0 and 1 the default value of 0.05 will be used).

3. Output files:

Sheet: Topological scoring (in and out) for (1) and (3):

Columns:

- 1). Entrez gene id of the gene from the shortest path
- 2). p value

p_out: statistical significance of the gene in shortest path (column 1) in providing connectivity (paths go from input set node towards shortest path node)

p_in: statistical significance of the gene in shortest path (column 1) providing connectivity (paths go from shortest path node towards input set node)

3). percentile: the “significant” shortest path genes are ranked. The percentile for a gene gives the percentage of significant genes falling below that significance level (higher p value).

4). the maximum number of genes from the input list that can be reached through the shortest path gene (column1) from another gene (the gene must be from the input list for (1); drug target for (2) and any gene for (3) and (4) algorithms). For directionality of paths see p_in and p_out.

5). drug targets of the gene if any

6). type of drug target interaction with the gene if any

Sheet: Topological scoring for (2) and (4):

For (2) and (4) all paths starts from a node and go towards a transcriptional factor target, therefore there is just one directionality (which would correspond to p in). The rest of the notation is similar to (1) and (3).

Sheet: ratio_out (in):

Its a matrix where the rows correspond to the genes from the input list and the columns to the genes from the shortest paths. The values in the matrices are calculated as follows: the number of times a node is part of the shortest path network when connecting all nodes from the network (for (3) and (4)) or the nodes from the input/drug target list (for (1) and (2)) to a node from the input list (which also must be

a transcription factor target for (2) and (4)), normalized by the number of nodes from where the same node can be reached via any shortest path.

Sheet : matrix_out (in):

The format is similar to ratio, but here instead of ratios we pick the best p value of those instances where the input node can be reached from the starting nodes via the given shortest path node.

For more questions about the algorithm **email:** zoltan@genego.com.