**Candidate miRNA regulatory hub identification pipeline:**

This pipeline is intended to compute the predicted impact for each miRNA in a set (typically those highly expressed in a particular cell type) on a network of genes (typically those relevant to the study of a disease relevant to that cell type), and to predict candidate miRNAs that may act as regulatory hubs in that cell type for a particular disease or pathway.

*Inputs:*

1. A list of predicted target sites for miRNA families in the multiple-sequence aligned 3’-UTRs of all genes for a set of species. This output can be generated from TargetScan (<http://www.targetscan.org>).
2. A list containing the conservation (number of species) of each of the miRNA families included in the above TargetScan predictions. This list was parsed out of the TargetScan input files.
3. A list of high confidence protein-protein interactions. This list was derived from the STRING 9.0 database (http://string-db.org/) using only those interactions having an interaction score greater than 700 (high-confidence). All protein identifiers were mapped to their corresponding gene symbol, and this symbol must match those used in the TargetScan predictions.
4. A list of miRNA families: This list contains the miRNAs the user wishes to include in the simulations. All miRNAs must be in the TargetScan miRNA family name format, and must match the family names in the Target Scan output file.
5. Gene list: A list of genes to use as central nodes in a gene network. Typically this list includes a set of genes relevant to the study of a particular disease or pathway. Currently all gene names must match the gene symbol used in the TargetScan 3’-UTR sequence files.

*Parameters:*

1. ***C***: Conservation level. Requested minimum level of conservation of each miRNA and target site required for a miRNA - gene interaction to be scored in the simulation.
2. ***N***: Number of iterations. Requested number of random gene networks of similar design that are used to generate score distributions.
3. ***α***: Hub weighting. This parameter is used to weight the contribution of the number of high-confidence protein-protein interactions to the target scoring function.

*miRNA scoring algorithm:*

A gene network is compiled using the input files for (1) the input gene list and (2) each of the ***N*** requested random gene networks. The input gene network contains all genes in the input gene list that have a 3’-UTR listed in the target prediction files, a weighted set of scores for each target site within each gene, and the number of high confidence protein - protein interactions listed for that gene in the STRING 9.0 database. Each random gene network is generated by selecting a set of random genes having connectivity similar to each of those in the input gene list. A gene is said to have similar connectivity if the gene has a similar number of high confidence interactions in the STRING 9.0 database. To compute groups of genes with similar connectivity, we group each gene in the STRING 9.0 database by the number of high confidence protein-protein interactions that gene has. If any group contains fewer than 20 genes, the group is expanded to include neighboring groups (with both higher and lower number of interactions) until the new super-group contains at least 20 genes. Finally a score is computed for each gene network (input gene network and ***N*** random gene networks) each miRNA family in the input list, and an empirical p-value is computed. The p-value is calculated as ***p=(Nr+1)/(N+1)***, where ***Nr*** is the number of random gene networks in which the targeting score for a particular miRNA was greater or equal to the score of that miRNA in the input gene network. The miRNA targeting score is calculated using the following procedure:

For a gene network ***G(L,D,U)****:* where ***L*** is the list of genes in the network, ***D*** is the number of high confidence protein-protein interactions that each gene has, and ***U*** is the ratio of the average 3’-UTR length in the input gene network over the average 3’-UTR length in the current gene network (note: this value is one when scoring the input gene network).

foreach ***miRi*** in ***miRlist***  having a conservation of at least ***C***:

foreach ***genej*** in ***L***:

foreach target site ***k*** of ***miRi*** in ***genej***:

foreach ***miRi*** in ***miRlist***  having a conservation of at least ***C***:

foreach ***genej*** in ***L***:

foreach additional target site ***k*** of ***miRi*** in ***genej***:

where *Posk* is the position of target site ***k*** within the 3’-UTR of ***genej***.