Configurability

Virtually all parameters used at every point in the execution of the program is configurable. The five categories of configuration are below; the number in parenthesis is the number of configurable options within that category.

1. General Configuration (8)
2. Analysis Configuration (8)
3. Force-Directed Layout Configuration (7)
4. Renderer Configuration (11)

Basic Configuration Options

General Configuration

- *activeDirectory* – directory from which all non-absolute paths will be resolved
  Default: <empty_string> (i.e. the directory the program is called from)
- *outputDirectory* – directory path for output to be saved
  Default: folder called "output" within *activeDirectory*
- *group1GeneSetFile* – file path that contains gene set information for Group 1
  Required
- *group2GeneSetFile* – file path that contains gene set information for Group 2
  Default: <empty_string>
- *projectName* – name of project; will be used as prefix for outputted file names
  Default: calculated based on *activeDirectory* and *group1GeneSetFile*

Analysis Configuration

- *minInteractomeConfidence* – the minimum STRING score (an integer between 0 and 1000) for a protein interaction to be considered in pairwise paths
  Default: 0
- *maxPathCost* - the maximum path cost for a pairwise path to be considered valid. A path consisting of interactions with confidence scores $c_1, c_2, \ldots, c_n$ has a path cost:

  $$cost = \sum_i 1000 - c_i$$

  Note that this path cost scheme the maximum confidence score to be 1000, as it is with default STRING downloads; otherwise Dijkstra's algorithm, as implemented, is not guaranteed to find the lowest cost path.
  Default: 200
- *maxPathLength* – the maximum number of vertices in a path. The maximum number of interactions, edges between source and target vertices, is $maxPathLength - 1$
  Default: 5

- *fractionOfVerticesToRender* – the fraction, between 0 and 1, of vertices to render in any graph
  Default: 1 (see **Note** below)
- *maxVerticesToRender* – the maximum integer number of vertices to render in any graph
  Default: 2147483647 (see **Note** below)
- *bootstrappingRounds* – the number of rounds of bootstrapping to perform. If specified, must be greater than 100. A value of 0 indicates no bootstrapping
  Default: 1000

Note: it is wise to set at least one of *fractionOfVerticesToRender* or *maxVerticesToRender*; otherwise, the program will attempt to render a large number of vertices, taking a long time to compute and most likely yielding uninterpretable results.

Renderer Configuration

- *displayRendering* – if false, images of the dendrogram and graphs will only be exported but not displayed to the user; this can provide a small speed improvement
  Default: true
- *significanceThreshold* – the p-value, between 0 and 1, at or below which clusters should be considered significant
  Default: 0.05
- *metaClusterThreshold* – the height, between 0 and 1, at which clusters should be grouped into meta-clusters; the higher the height, the larger the metaclusters
  Default: 0.3333

File Formats

The input gene list group file format is a series of individual patient gene sets, one per line:
<patient_id1>=<gene_symbol>,<gene_symbol>,<gene_symbol>,…,<gene_symbol>
<patient_id2>=<gene_symbol>,<gene_symbol>,<gene_symbol>,…,<gene_symbol>
*Please see examples for input gene set group files.

Exported files are either standard image files, .csv files, simple .txt tab-delimited tables, or .txt Newick tree files.

The Newick tree file format is as follows:

$$\big((\text{leafA}: 0, \text{leafB}: 0): \text{clusterABheight}, (\text{leafC}: 0, \text{leafD}: 0): \text{clusterCDheight}\big): \text{clusterABCDheight}$$

and can be arbitrarily nested. Programs that expect dendrogram files will recognize this format.

Advanced Configuration Options

Advanced Configuration Options (most users will **not** need to modify these)

General Configuration

- *reusePreviousData* – if set to false, will force the program to discard any previous pairwise path data
  Default: true
- *calculateGraphDifferences*– whether or not graph differences (Group1 – Group2 and Group2 – Group1) should be calculated, rendered, displayed, and exported
  Default: true
- *proteinInteractomeFile* – an **absolute** file path to a downloaded STRING protein interactome; if the file is compressed with GZIP, it must have the .gz extension
  Default: installation directory + STRING download's file name
- *proteinAliasesFile* – an **absolute** file path to a downloaded STRING protein aliases list; if the file is compressed with GZIP, it must have the .gz extension
  Default: installation directory + STRING download's file name

Force-Directed Layout Configuration

- *repulsionConstant* – the repulsion constant λ_R used in force-directed layout equations
  Default: 0.2
- *attractionConstant* – the attraction constant λ_A used in force-directed layout equations
  Default: 0.0003
- *minVertexRadius* – the minimum radius of vertices in the graphs being rendered
  Default: 15
- *maxVertexRadius* – the maximum radius of vertices in the graphs being rendered or -1 for dynamically determined maximum
  Default: -1

Stopping Conditions for Force-Directed Layout algorithm

- *deltaThreshold* – stop when the total absolute change summed over all vertices is less than or equal to *deltaThreshold*
  Default: 0.001
- *maxIterations* – the maximum number of iterations to update vertex positions
  Default: 10000
- *maxTime* – the maximum number of milliseconds to spend laying out the vertices
  Default: 9223372036854775807 (maximum 64-bit signed integer)

Renderer Configuration

- *minVertexAlpha* – the minimum alpha value (between 0 and 255) of vertices in the graph
  Default: 50
- *minEdgeAlpha* – the minimum alpha value (between 0 and 255) of edges in the graph
  Default: 50
- *drawGeneSymbols* – boolean flag indicating if gene symbols should be written on vertices
  Default: true
- *colorSignificantBranchLabels* – boolean flag indicating if cluster identifiers should be colored red if deemed significant (true), or should always be black (false)
  Default: true
- *defaultVertexColor* – the default color for vertices in a graph
  Default: (255,0,0) **(red)**
- *group1VertexColor* – the color for vertices whose gene is in from a Group 1 gene set
  Default: (255,200,0) **(yellow)**
- *group2VertexColor* – the color for vertices whose gene is in from a Group 2 gene set
  Default: (0,0,255) **(blue)**
- *bothGroupsVertexColor* – the color for vertices whose gene is in both Group 1 and 2 gene sets
  Default: (0,255,0) **(green)**

Color Format: Color configuration options are given by "(R,G,B)" where each of R, G, and B are replaced with an integer between 0 and 255 for red, green, and blue components

DEFAULT CONFIGURATION OPTIONS

To generate the below configuration file, you can run Proteinarium with the −d flag.

```
# Analysis Config
reusePreviousData            = true
calculateGraphDifferences    = true
minInteractomeConfidence     = 0.0
maxPathCost                  = 200.0
maxPathLength                = 5
fractionOfVerticesToRender   = 1.0
maxVerticesToRender          = 2147483647
bootstrappingRounds          = 1000

# Force Directed Layout Config
repulsionConstant            = 0.2
attractionConstant           = 3.0E-4
deltaThreshold               = 0.001
maxIterations                = 10000
maxTime                      = 9223372036854775807
minVertexRadius              = 15.0
maxVertexRadius              = -1.0

# General Config
activeDirectory              =
outputDirectory              = output/
group1GeneSetFile            = <no default>
group2GeneSetFile            =
projectName                  = <group1GeneSetFile>
proteinInteractomeFile       = <path>/9606.protein.links.v11.0.txt.gz
proteinAliasesFile           = <path>/9606.protein.aliases.v11.0.txt.gz
stringDatabaseVersion        = 11.0

# Renderer Config
displayRendering             = true
minVertexAlpha               = 50
minEdgeAlpha                 = 50
drawGeneSymbols              = true
defaultVertexColor           = (255,0,0)
group1VertexColor            = (255,200,0)
group2VertexColor            = (0,0,255)
bothGroupsVertexColor        = (0,255,0)
colorSignificantBranchLabels = true
significanceThreshold        = 0.05
metaClusterThreshold         = 0.3333
```

Sample Configuration File

group1GeneSetFile is required.

If you wish to run analyses or visualize two groups at once, ***group2GeneSetFile*** is required.

If no ***projectName*** is specified, it will be the name of the file provided in ***group1GeneSetFile***

All options not specified in the configuration will take their values from the default configuration options shown above.

config.txt example file:

```
group1GeneSetFile=SIMdataset1_original.txt
group2GeneSetFile= SIMdataset2_original.txt
projectName=SIMset_results
maxVerticesToRender=50
metaClusterThreshold=0.8
maxPathLength=5
```