Manual and algorithm of SUbNetworks RIch in Significant Elements (sunrise) in C++

1 introduction

sunrise searches subnet significantly enriched in low p-values. It needs a network as a protein protein interaction network and an additive score which can be computed from p-values obtained by genome-wide association study or gene expression. The highest score subnet gives information about the pathways or groups of proteins influenced by the score.

A score is associated with every node, edge weights are not used. Scores must be additive. Groups of nodes are simply described as an array of boolean and the below genetic algorithm is used. The below algorithm searches the highest score subnet (it is one of the Steiner tree problems). The undirected network is connected without loop and duplicated edges.

2 Use

Written in C++ 11, the sources must be compiled in your system.

The executable is launched with the name of parameters file in the command line.

The better result is displayed by seed and for all seeds. The lists of nodes of the network and of the reduced network are saved.

The result is sensitive to below parameters notably population and mutation.
Empirically, star parameters may be:
population = number of negative nodes / 2,
generation = population * 10,
mutation = population / 2,
seed = several values
The best results and the corresponding parameters can be kept as comments in the parameter file.

3 Format of files

Files are made from the name of the project project_name given in the parameter file. All files can be loaded in Cytoscape as network or attributes. The result is a simple list of nodes for selecting nodes by ID in Cytoscape. Lines must end by $< CR >< LF >$ or <\r\n> and fields separated by $< tab >$.

3.1 Input files

- network: project_name.sif
  node $< tab >$ *interaction* $< tab >$ *node*, without header

- nodes: project_name.txt
  node $<$ *tab* $>$ *score*, with header *shared name* $<$ *tab* $>$ *score*

3.2 intermediate files

- reduced network: project_name-r.sif
  group of nodes $<$ *tab* $>$ *pp* $<$ *tab* $>$ *group of nodes*, without header
- attributes of groups of nodes: project_name-a.txt
  group of nodes $<$ *tab* $>$ *score of groups* $<$ *tab* $>$ *list of nodes in the group*, with header *shared name* $<$ *tab* $>$ *score* $<$ *tab* $>$ *name_in_group*

3.3 output files

- result groups: project_name-p.txt, list of groups of nodes
- result nodes: project_name-o.txt, list of nodes

4 parameter file

The parameters are, order does not matter:
Several lines of comments beginning by #
dir = work directory
project = name of project used to name all files
step = 0 or 1 or 2
0: reducing and searching, 1: reducing only, 2: searching only
seeds = seeds separated by semicolon
population = size of a population
generation = number of generation
mutation = number of mutations

5 Algorithm

When loading a network, nodes of network and reduced network are sorted by score.

5.1 Reduce Network

A reduced network containing the optimal subnet is built in two stages.

5.1.1 Group Positive Nodes

All connected positive nodes are gathered in groups. When rebuilding the network, parallel edges and auto-loops are eliminated. Scores of absent nodes in network are set to zero and treated as a positive node.

5.1.2 Keep Shortest Paths

Connecting the positive groups needs to pass by negative nodes. Edges are doubled and negative scores of ends are transferred to edges with reversing edge signs. Only the shortest paths between positive nodes in a graph are kept.

5.2 Genetic Algorithm

5.2.1 Start

The pseudo-chromosomes are described as an array of booleans. The highest score shortest paths between positive groups are taken as the start population. All positive nodes are set to true.

5.3 Function To Maximize

The function to maximize take the group of connected nodes having the maximal score.The useless nodes of the group are not taken into account to compute the sum of scores (one degree negative nodes are recursively removed and inside negative nodes not changing the number of linked positive nodes are removed in increasing order). The pseudo-chromosome is not modified.

5.3.1 Search by Genetic Algorithm

Every seed launches a random sequence of integers. This sequence is used for mating, cross-over and mutations. Only items having the highest score by function to maximize are kept in population and so on until the number of generations is reached.

October 30, 2023 by daniel.rovera@gmail.com - Institut Curie