UNIVERSIDADE FEDERAL DE JUIZ DE FORA

Programa de Pós Graduação em Modelagem Computacional

CGPGRN - Change Log

José Eduardo Henriques da Silva

February 14, 2023

Contents

1 Introduction

CGPGRN is a framework for inferring Gene Regulatory Networks (GRNs) using time-series data which is the result of a Ph.D. Thesis. Initially developed for single-cell data, CGPGRN can be used for any data-type technology.

This change log contains information about the modifications in CGPGRN framework.

More information about CGPGRN framework can be found at

`https://github.com/jeduaardo/cgpgrn`.

2 Change Log

2.1 Version 1.0

- Initial release
  - Base functionalities for CGPGRN framework: pre-processing, clustering, discretization, inference and post-process.

2.2 Version 2.0

- Multiple pseudotimes support;
- Spline Files can be passed as argument;
  - for a single file, the argument is the file itself
  - for multiple spline files, -sp must contain a .txt file with one spline file per line
- Possibility of no cluster method usage (-cm 'None', default);
- Log file with the configuration used for preProcessing data (CGPGRN_parameters.txt);
- postProcess for generating rankedEdges and organizing the source data used. It also make the root directory clean;
- Automatic execution of postProcess script;
- include folder added with all scripts used with CGPGRN;

- added requirements folder with getRequirements script for getting/updating the CGPGRN's required python libraries
- Linux support - issues fixed

2.3 Version 2.1

- Multiprocessing support (more than 10 speed up);
- Agglomerative Clustering added in clustering methods;
- Minor Optimizations (utils, discretization and clusterMethods);
- Unified directories for clustering methods and discretization prefixes in utils;
- Fixes on BiKMeans discretization procedure;
- Added -run argument for choosing running or not CGPGRN inference algorithm after the data processing;
- Added time counting in each CGPGRN framework step. Stored in CGPGRN_times.txt;
- Added LogFile with information about the CGPGRN framework pipeline execution;
  - Added respective Log object
- All arguments are now verified in utils;
- Minor Bugfixes:
  - When using full discretization, when nc < n_genes - 1 now the framework returns a notification and automatically uses not full discretization. All information is presented in prompt command and in Log file;
  - Fixed the bits conversion in ambiguous transitions in discretization.
- Added -kd argument for keeping data when the framework pipeline is done. This keeps the source and spline data on the CGPGRN framework main directory, allowing for new executions;

2.4 Version 2.2

- New general requirements: G++ Compiler, Make and Bash.
- added support for multiple outputs and don't care recognizing;
- -kd default changed to False
- -run default changed to False
- added -gsl argument for automatically generating spline list files (default = False)
- added makeFile.py in includes for automatically generating Makefile
- Added CGPGRN inference algorithm parametrization:
  - -cgpn defines the number of nodes
  - -cgpg defines the number of generations
- added timeStatistics for counting time in the entire framework
  - inference (mean and total)
  - complete framework (total)
- Minor fixes and more information on Log
- added mkdir function in utils
- all functions are now documented in utils
- noClustering now can be used with -fullTT
- fullBashScript is now generated in utils
  - automatically defines the number of nodes and generations of CGPGRN inference algorithm. However, the user still can define
- objects added for performing the CGPGRN framework main steps
  - spline
  - clustering
  - discretization

- rankedEdges now generates one unified rankedEdges containing all independent runs and one rankedEdges for each independent run