User Manual of CLAM

Contents

1 Introduction

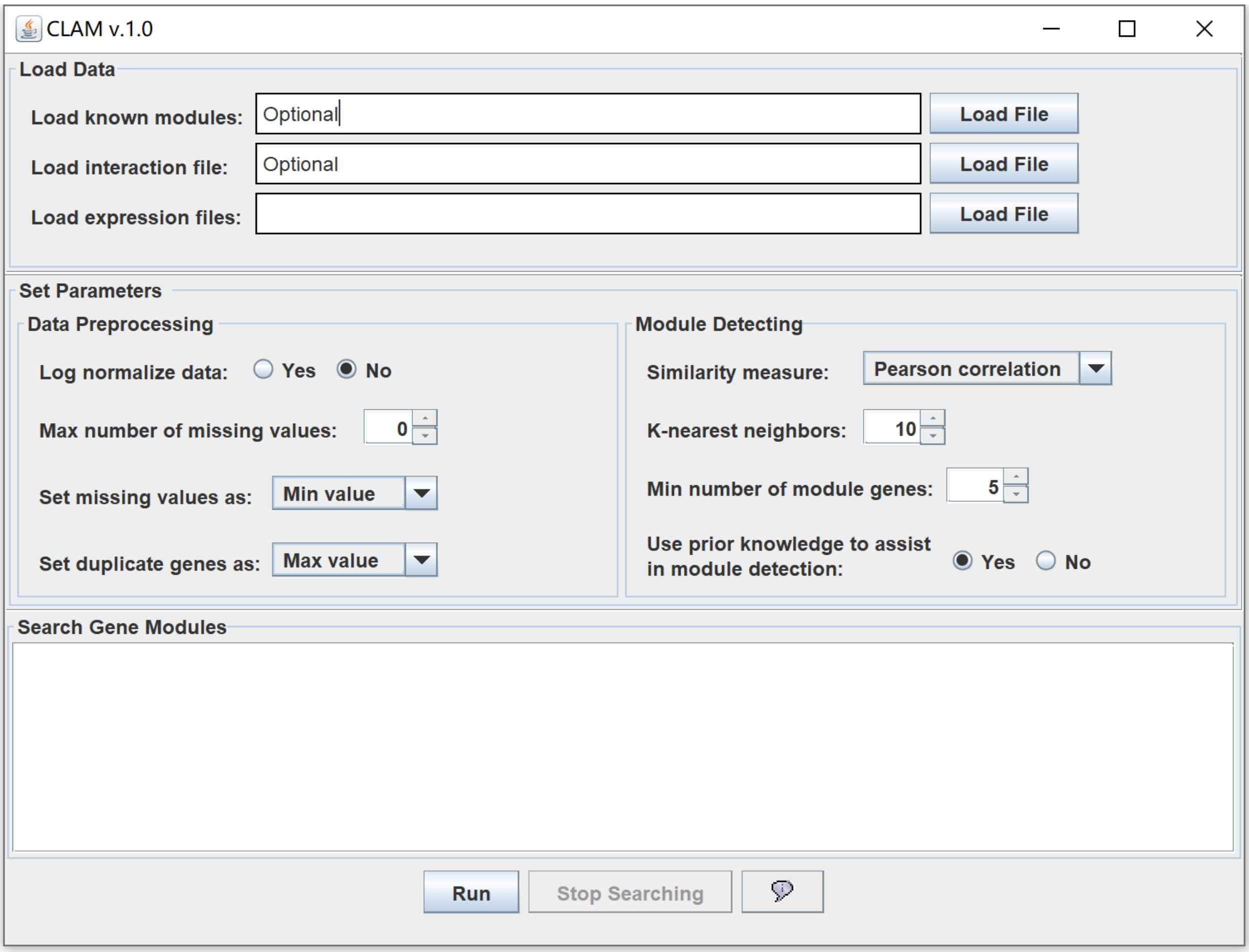
CLAM is an analytical framework for identifying co-expressed gene modules by integrating multi-omics data and known molecular interactions. The source code and the complied tool can be downloaded at https://github.com/free1234hm/CLAM.

2 Preliminaries

- To use CLAM a version of Java 1.5 or later must be installed. If Java 1.5 or later is not currently installed, then it can be downloaded from http://www.java.com.
- CLAM can be executed by double-clicking on ‘CLAM.jar’, or from a command line change to the CLAM directory and then type: java -mx1024M -jar CLAM.jar. If CLAM reports ‘Out of Memory Error’, users can increase the -mx parameter.

3 Main Interface

The main interface has three parts: ‘Load Data’, ‘Set Parameters’, and ‘Search Gene Modules’.


CLAM v.1.0

Load Data

Load known modules: Optional

Load interaction file: Optional

Load expression files:

Set Parameters

Data Preprocessing

Log normalize data: ○ Yes ● No

Max number of missing values: 0

Set missing values as: Min value

Set duplicate genes as: Max value

Module Detecting

Similarity measure: Pearson correlation

K-nearest neighbors: 10

Min number of module genes: 5

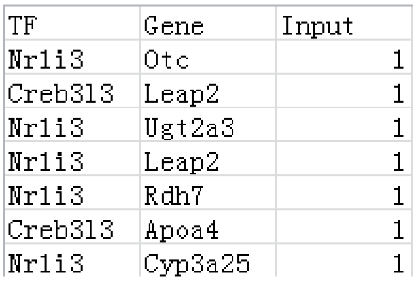
Use prior knowledge to assist in module detection: ● Yes ○ No

Search Gene Modules

Figure 1. The main interface of CLAM

3.1 Load known modules

Import known module files, such as GO annotation, KEGG pathway information, transcription factor (TF)– and microRNA (miRNA)–target interaction information. The source of these information can either be user provided or one of the files present in the 'Known modules' directory. The known module files should be in a three-column format: the first column contains module name, the second column the regulated gene, and the third column input value. The first row is a header row where the header of the first column can be 'TF', 'miRNA' and 'pathway' et al, and the second column must have the header 'Gene'. A value of '1' represents that the TF-gene pair shows positive correlation interaction, while a value of '-1' represents that the TF-gene pair shows negative correlation interaction (Figure 2).



TF	Gene	Input
Nr1i3	Otc	1
Creb3l3	Leap2	1
Nr1i3	Ugt2a3	1
Nr1i3	Leap2	1
Nr1i3	Rdh7	1
Creb3l3	Apoa4	1
Nr1i3	Cyp3a25	1

Figure 2. A sample of TF-gene data file in three-column format when viewed in Microsoft Excel.

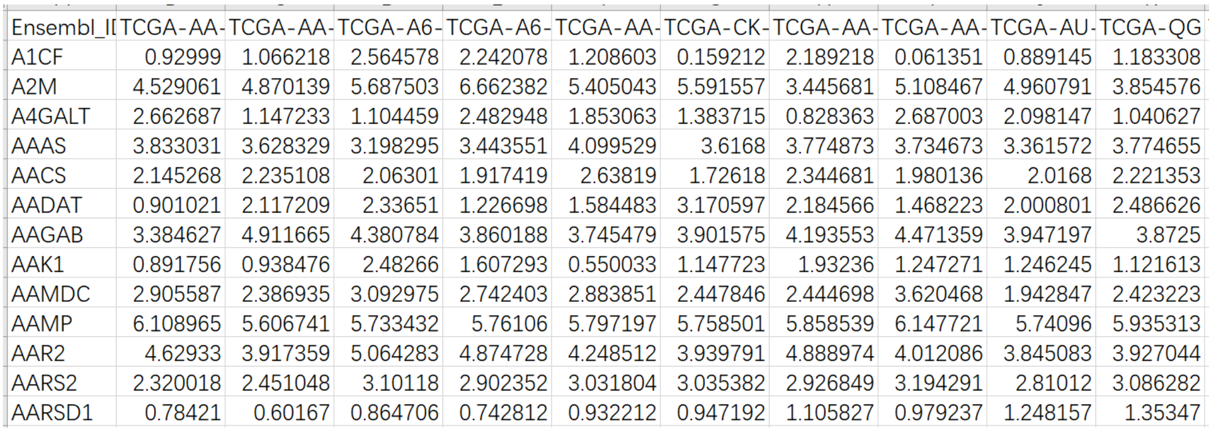
The known modules can be used as prior knowledge to improve module detection (only when 'Use prior knowledge to assist in module detection' is selected), and used in functional enrichment analysis after module detection.

3.2 Load interaction file

Import protein–protein interaction (PPI) data, which can either be user provided or one of the files present in the 'Interaction files' directory. Same as known modules, the PPI data should be in a three-column format. The PPI interactions can be used as prior knowledge to improve module detection (only when 'Use prior knowledge to assist in module detection' is selected).

3.3 Load expression files

The 'Load expression files' field is used to import datasets from different sources. All datasets need to written in same format: the first column is gene or protein names, and the remaining columns contain the measurements in different samples. If a measurement is missing, then the field should be left empty. The first row of the data contains column headers. **Notably, CLAM does not require different omics datasets to share the same genes (proteins) or samples.** A sample expression data file is shown in Figure 3.



Ensembl_II	TCGA-AA-	TCGA-AA-	TCGA-A6-	TCGA-A6-	TCGA-AA-	TCGA-CK-	TCGA-AA-	TCGA-AA-	TCGA-AU-	TCGA-QG
A1CF	0.92999	1.066218	2.564578	2.242078	1.208603	0.159212	2.189218	0.061351	0.889145	1.183308
A2M	4.529061	4.870139	5.687503	6.662382	5.405043	5.591557	3.445681	5.108467	4.960791	3.854576
A4GALT	2.662687	1.147233	1.104459	2.482948	1.853063	1.383715	0.828363	2.687003	2.098147	1.040627
AAAS	3.833031	3.628329	3.198295	3.443551	4.099529	3.6168	3.774873	3.734673	3.361572	3.774655
AACS	2.145268	2.235108	2.06301	1.917419	2.63819	1.72618	2.344681	1.980136	2.0168	2.221353
AADAT	0.901021	2.117209	2.33651	1.226698	1.584483	3.170597	2.184566	1.468223	2.000801	2.486626
AAGAB	3.384627	4.911665	4.380784	3.860188	3.745479	3.901575	4.193553	4.471359	3.947197	3.8725
AAK1	0.891756	0.938476	2.48266	1.607293	0.550033	1.147723	1.93236	1.247271	1.246245	1.121613
AAMDC	2.905587	2.386935	3.092975	2.742403	2.883851	2.447846	2.444698	3.620468	1.942847	2.423223
AAMP	6.108965	5.606741	5.733432	5.76106	5.797197	5.758501	5.858539	6.147721	5.74096	5.935313
AAR2	4.62933	3.917359	5.064283	4.874728	4.248512	3.939791	4.888974	4.012086	3.845083	3.927044
AARS2	2.320018	2.451048	3.10118	2.902352	3.031804	3.035382	2.926849	3.194291	2.81012	3.086282
AARSD1	0.78421	0.60167	0.864706	0.742812	0.932212	0.947192	1.105827	0.979237	1.248157	1.35347

Figure 3. A sample of expression data file when viewed in Microsoft Excel.

3.4 Set Parameters

Through the parameters on the 'Data Preprocessing' panel the user can adjust the criteria for filtering genes. If a gene is filtered, then it will be excluded from further analysis. Assuming that the expression vector of a gene is $\{v_1, v_2, \dots, v_n\}$.

- Log normalize data—transforms the vector to $\{\log_2 v_1, \dots, \log_2 v_n\}$.
- Maximum number of missing values—a gene will be filtered if the number of missing values in all samples exceeds this parameter.
- Set missing value as—the missing value of a gene can be set as 'min value' (default), 'mean value' or 'zero'.
- Set duplicated genes as—combine the duplicate expression profiles of a gene based on the

‘mean value’ or ‘max value’ (default).

- Similarity measure—the measure to estimate the similarity between objects (genes or proteins), including Euclidean distance, Pearson correlation coefficient, absolute Pearson correlation coefficient, cosine similarity and mutual information. The absolute Pearson correlation coefficient and mutual information can assign objects with similar and opposite features to the same module.
- K-nearest neighbors—search the $\boldsymbol{k}$ nearest neighbors for each gene. The higher $\boldsymbol{k}$ is, the fewer cluster centers will be identified; as a consequence, fewer clusters will be generated.
- Min number of module genes—the minimum (5 as default) number of genes in a module.
- Use prior knowledge to assist in module detection—if ‘Yes’ is selected, the files that currently is present in the ‘Interaction data’ directory of the CLAM directory will be used to assist in module detection.

3.5 Search Gene Modules

The text box here displays the running progress of CLAM after pressing the ‘Run’ button. After the analysis process is finished, CLAM will create a new folder named ‘Results’ in its directory to save the analysis results, including the table of final modules and the functional enrichment analysis results.