IMPACT-modules

Documentation of the software performing the network-based analysis

This method combines the phenotypic data (i.e. RNAi screening data, etc…) with binary interaction data, such as physical protein binding data. In this case the aim is to screen the network for sub-networks (modules) enriched for a common phenotypic signature. The components of such network modules are assumed to be involved in the same or related pathways or biological processes.

Main function "IMPACT_modules"

Call from inside MATLAB: **IMPACT_modules (do_params, alg_params)**
where **do_params** = [par1 par2 par3 par4] and **alg_params** = [par5, par6, par7, par8]

Example call: **IMPACT_modules ([1 1 1 1], [0.7 2 5000 4])**

Short parameters description

INPUT:

do_params = [do_seed_nodes do_module_search do_randomizations do_statistics]

par1: do_seed_nodes = binary, [0,1], do the seed node finding
par2: do_module_search = binary, [0,1], do the search procedure on network
par3: do_randomizations = binary, [0,1], do the randomizations
par4: do_statistics = binary, [0,1], calculate p-values and export results based on the randomization results

alg_params = [SIMIL_THR MIN_NUMB_PROF N_RAND N_MAX]

par5: SIMIL_THR = number, [0-1], similarity threshold (Pearson correlation coefficient) for search and randomizations
par6: MIN_NUMB_PROF = number, if >= 1 it is the number of profiles above similarity threshold required for module search (e.g. 2); if < 1 it is the fraction of profiles above threshold.
Par7: N_RAND = integer number, e.g. 1000, number of randomization for each set
Par8: N_MAX = integer number, e.g. 4, number of bins on the number-of-profiles distribution for the randomizations

OUTPUT: .mat and .txt files in the same folder as input files

example call IMPACT_modules([1 1 1 1], [0.7 2 5000 4])

Long parameters description

Environment parameters

1. ***par1*** (import and mapping)
    - Value = 1 -- importing both interaction and phenotypic data from text files: a modal dialog box window will appear for each of the two files so that the user can choose and load first the file containing set interaction data and then the file containing the phenotypic data. The software saves the mapping information in a .mat file that can be loaded in the subsequent steps of the analysis.

      After mapping,, an additional step is performed for estimating the starting seed nodes for module expansion.
    - Value = 0 – this step is skipped: the software assumes that in this case the import and mapping were performed in advance. Useful when re-running the searching procedure on previously imported data.

2. ***par2*** (pattern search)
    - Value = 1 -- performing the searching procedure and saving the results in an intermediate .mat file.
    - Value = 0 -- this step is skipped: the software assumes that in this case the search for enriched pattern was performed in advance. Useful when re-running the statistical assessment (randomizations) on network modules previously detected.

3. ***par3*** (randomizations)
    - Value = 1 -- performing the randomization procedure necessary for the estimation of p-values associated to each of the set analysed. An intermediate .mat file is created.
    - Value = 0 -- this step is skipped: the software assumes that in this case the randomization was previously performed. Useful when re-running only the last step, i.e. p-value calculation and export of final results.

4. ***par4*** (p-value estimation and results export)
    - Value = 1 -- loading data from the previous steps of analysis and computing the p-values. The final results will be stored in a .mat file and also exported in a .txt file in the form of a table where columns are tab separated.
    - Value = 0 -- this step is skipped.

Algorithm parameters

5. ***par5*** (similarity threshold)
    - Similarity threshold in terms of Pearson correlation coefficient that will be used for the searching procedure. It must be a numeric value between 0 and 1.
6. ***par6*** (minimal number of similar profiles requested)
    - Minimal number of similar profiles per each gene (i.e. having correlation higher than the similarity threshold) requested to inclusion during module expansion in the network-based approach. If integer, it represents the actual minimal number of profiles; if fractional number, it represents the minima fraction of similar profiles requested to be similar. For example, if set to 0.5, it means that 50% of the profiles of a certain gene should be more similar than the requested similarity threshold for allowing gene inclusion in the expanding network module.
7. ***par7*** (number of randomization)
    - Number of randomization that the software will perform. It can be any integer value. The randomizations are performed for the estimation of the empirical probabilities of similarity associated to the pattern profiles used during the searching step.
8. ***Par8*** (number of bins)
    - Maximal number of bins that the software will define for the sampling step in the randomization procedure. It is used to perform randomization stratified on the number of profiles per gene, to avoid biases (i.e., genes with a higher number of profiles have different statistical power than genes with lower one). This parameter should be set to a reasonable integer (e.g., n = 2, 3 or 4) value that will divide the total dataset in more or less n equally sized group still having sufficient number of elements. Set 1 if no stratification is desired.

INPUT FILES

In order to perform either the set- or the network-based analysis it is necessary to load 2 text file (.txt):

1. **Interaction network file** (for the network-based analysis);
2. **Phenotypic data file**.

See paragraphs below for detailed explanation.

Interaction network file

The software expects a .txt file containing an interaction network stored in a table where the columns are tab-delimited, as in the following example:

Node_1	**Node_2**
GeneA	GeneB
GeneA	GeneC
…	…

Note: the identifiers of the network nodes must match the identifiers of the elements present in the phenotypic data file in order to perform the mapping between the two datasets.

Phenotypic data file

The software expects a .txt file containing the phenotypic data in tab-separated columns, as in the following example:

Genes ID	**Profile ID**	**Param1**	**Param2**	**Param3**	**…**
GeneID_1	GeneID1_Prof1	value_param1	value_param2	value_param3	…
GeneID_1	GeneID1_Prof2	value_param1	value_param2	value_param3	…
GeneID_2	GeneID2_Prof1	value_param1	value_param2	value_param3	…
…	…	…	…	…	…

Note: the first column has to contain the identifiers of the elements under investigation. These identifiers must match the identifier present in the phenotypic data file in order to perform the mapping between the two datasets.

OUTPUT FILES – NETWORK-BASED ANALYSIS

The software automatically saves the results of the analysis in a .txt files named:

NamePhenotypicInputFile_output_export-100thresholdValue.txt

Where:

- *NamePhenotypicInputFile* is the filename of the file containing the phenotypic data
- *100thresholdValue* is a suffix added at the end of the filename and it consists of the number obtained by multiply the similarity threshold by 100.

The file will contain the results in a table where columns are tab-separated.

NUM_MODULE	p-value	#tot_genes	#tot_profiles	Gene_ID	Profiles_Symbols
1	Value	Value	Value	GeneID1, GeneID2, …	GeneID1_Prof1, GeneID1_Prof2, …
2	Value	Value	Value	…	…
…	…	…	…	…	…

Note: the 5th and 6th columns contains respectively the identifiers of the genes selected in the set and the identifiers of the profiles selected separated by commas.

The software produces also several files that contain the results of intermediate steps of the analysis. Those files are listed below:

- *NamePhenotypicInputFile_pheno-mapped.mat*

It contains the interaction network and phenotypic data after the mapping step;

- *NamePhenotypicInputFile*_seed-file_prof-*100thresholdValue*.mat and *NamePhenotypicInputFile*_seed-list_prof-*100thresholdValue*.txt

These files contain the information about the seed to use for starting the module expansion step and their relative phenotypic profiles.

- *NamePhenotypicInputFile*_res-search-file_prof-*100thresholdValue*.mat

It contains the results of the searching procedure (i.e. information of the modules and of their relative selected profiles and reference profiles);

- *NamePhenotypicInputFile_p-distr-file_prof-100thresholdValue.mat*

It contains the probability distributions (analytically computed) relative to each of the seed pattern used for the searching step.

- *NamePhenotypicInputFile*_res-stat-file_prof-*100thresholdValue*.mat

It contains the final results of the network-based analysis i.e. set, their relative selected profiles and reference profiles, p-values.