IMPACT-sets

Documentation of the software performing the set-based analysis

The set-based method analyses sets of genes with common properties (e.g. genes encoding components of a multi-protein complex or common pathway). It identifies an enriched phenotype that is shared by a maximum number of genes/proteins in the set. We achieve this goal using an approach that is similar to common cluster analysis, but constrained on the members of a given set: the algorithm identifies the largest group (cluster) of profiles spanning a maximum number of genes in the set.

Main function "IMPACT_sets"

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

Example call: **IMPACT_sets ([1 1 1 1], [0.75 1 5000 4])**

Short parameters description

INPUT:

do_params = [do_import_mapping do_set_search do_randomizations do_statistics]

par1: do_import_mapping = binary, [0,1], do the import (convert input text files into matlab environment files)
par2: do_set_search = binary, [0,1], do the search procedure on sets
par3: do_randomizations = binary, [0,1], do the randomations
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 = not used, it must be passed to the function (it may be used in future implementations);
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_sets([1 1 1 1], [0.75 1 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.
    - 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 sets 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. A randomization folder is created and an intermediate .mat file is created in there.
    - 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 the search for enriched pattern. This value is ignored in the current implementation, but it still must be passed to the function for compatibility reasons with the network approach. It may be used in future implementations.
7. ***par7*** (number of randomization)
    - Number of randomization that the software will perform. It can be any integer value. Randomization is repeated the number of times specified by this parameter for each set, to empirically estimate of the p-values.
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. **Set information file** (for the set-based analysis);
2. **Phenotypic data file**.

See paragraphs below for detailed explanation.

Set information file

The software expects to read a .txt file containing the set information in a table where the columns are tab-delimited, as in the following example:

Set_database_ID	Set_Name	Set_Components	Set_Organism	Set_db_id
SET1	Name_Set1	GeneA,GeneB,GeneC	Human	Set_dbID1
SET2	Name_Set2	GeneE,GeneF	Human	Set_dbID2
…	…	…	…	…

Note: the third column has to contain the identifiers of the components of each set separated by commas. These identifiers must match the identifier 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 – SET-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_SET	Set_Name	p-value (Gene)	p-value (Profiles)	#tot_genes_sel	#tot_profiles_sel	geneID_sel	profID_Sel
SET1	Name_Set1	Value	Value	Value	Value	GeneID_1, GeneID_2, GeneID_3	GeneID1_Prof1, GeneID1_Prof2, GeneID2_Prof1 GeneID3_Prof1
SET2	Name_Set2	…	…	…	…	…	…
…	…	…	…	…	…	…	…

Note: the 7th and 8th 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*

contains set information and phenotypic data after the mapping step;

- *NamePhenotypicInputFile*res-search_thr-*100thresholdValue*.mat

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

- *NamePhenotypicInputFile*_res-final_thr-*100thresholdValue*.mat

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