Documentation for IMPACT

IMPACT (Integrative Multi Profile Analysis of Cellular Traits) performs integration of functional screens and molecular interaction data.

The main goal of IMPACT is to identify the most consistent phenotypic profile among interacting genes. This approach utilizes two types of external information: sets of related genes (IMPACT_sets) and network information (IMPACT_modules). Based on the notion that interacting genes are more likely to be involved in similar functions that non-interacting genes, this data is used as a prior to inform the filtering of phenotypic profiles that are similar among interacting genes. IMPACT_sets selects the most frequent profile among a set of related genes. IMPACT_modules identifies sub-networks containing genes with similar phenotype profiles. The statistical significance of these selections is subsequently quantified via permutations of the data.

STANDALONE APPLICATION

It is possible to download the standalone applications that perform the set- and network-based analysis from the following website:

http://cellnet.cecad.uni-koeln.de/impact.html.

In the current implementation, only the standalone application for Windows operating systems is available (.exe).

The source code can be found at the same location, and standalone application for other platforms can be compiled by using the respective versions of the MATLAB Compiler ™.

In this document there is a short description of the input files that the software can read and the parameters that the user can set for performing the Set-based and the Network-based analysis.

SHORT DESCRIPTION OF THE METHODS

IMPACT_sets: set-bases 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.

IMPACT_modules: network-based analysis

The network-based 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.

INPUT FILES

In order to perform either the set- or the network-based analysis it is necessary to load 3 text file (.txt):

1. A **parameter file**;
2. **Set information file** (for the set-based analysis) or **Interaction network file** (for the network-based analysis);
3. **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.

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.

Parameter file

The software expects a .txt file containing 8 parameters for the method to run, all in one line (either comma or tab or space separated), as in the following example:

par1, par2, par3, par4, par5, par6, par7, par8

See description below for extensive explanation. Please note that the order of the parameters is crucial.

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.
      For the network-based analysis, 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 sets/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. For sets, a randomization folder is created and an intermediate .mat file is created in there. For network modules, an intermediate .mat file only 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.

This value is ignored in the set-based analysis, but it still must be specified in the parameter file for compatibility reasons with the network approach.

7. ***par7*** (number of randomization)
    - Number of randomization that the software will perform. It can be any integer value. For the set-based analysis, the randomization step is performed for the empirical estimation of the p-values. For the network-based analysis, the randomizations step is 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.

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.

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.

HOW TO RUN THE SOFTWARE TOOL

Download the software from the website
http://cellnet.cecad.uni-koeln.de/impact.html
and decompress the zip file in any destination folder. Then, follow the instructions below.

Standalone distribution

Windows Installer

IMPACT-sets: before running the application, run the setup file `IMPACT_sets_Installer_mcr.exe` (inclusive of the MATLAB Compiler Runtime (MCR) libraries).
After installation run `IMPACT_sets.exe from the installation folder.`

IMPACT-modules: before running the application, run the setup file `IMPACT_modules_Installer_mcr.exe` (inclusive of the MATLAB Compiler Runtime (MCR) libraries).
After installation run `IMPACT_modules.exe from the installation folder.`

EXAMPLE1

Run `IMPACT_sets.exe`

`with parameters 1, 1, 1, 1, 0.7, 1, 5000, 4 (stored in input_params_file.txt).`

This instruction perform:

- import/mapping of the data;
- set-based search for each set considering 0.7 as similarity threshold;
- 5000 randomizations for each set;
- creation of 4 bins of the mapped phenotypic data: those are used in the randomizations step.

EXAMPLE2

Run `IMPACT_modules.exe`

`with parameters 1, 1, 1, 1, 0.7, 2, 5000, 4 (stored in input_params_file.txt).`

This instruction performs:

- import/mapping of the data and seed nodes definition;
- network-based analysis using 0.7 as similarity threshold and 2 as minimal number of similar profiles;
- 5000 randomizations for the empirical estimation of the similarity probability associated to the pattern used for the search;

- creation of 4 bins of the mapped phenotypic data: those are used in the randomizations step.