OptiCon

version 1.1
Dec 12th, 2018

Overview

Complex diseases are caused by multiple deregulated pathways. To achieve an effective and lasting treatment, multiple pathways need to be targeted in combination. Current combination therapies are developed mainly based on targets of existing drugs, which only represent a small portion of the human proteome. For *de novo* identification of combinatorial therapeutic targets, systematic identification of synergistic key regulators of disease conditions represents an effective strategy. We introduce a network controllability-based method, ***OptiCon***, to identify synergistic regulators as candidate targets for combination therapy. These regulators jointly exert maximal control over deregulated genes but minimal control over unperturbed genes in a disease. Therefore, modulating these regulators maximizes the likelihood of correcting gene deregulation and minimizes the likelihood of side effect. Using data from three types of cancer, we show that 69% of predicted regulators are either known drug targets or have a critical role in the development of a given cancer type. The predicted regulators are also depleted for known proteins associated with side effect. The predicted synergy is further supported by the enrichment of recurrently mutated cancer genes, disease-specific and/or clinically relevant synthetic lethal interactions, functional gene interactions among the co-regulated subnetworks, and experimental validation. Moreover, we find that a significant portion of genes regulated by synergistic regulators are involved in dense interactions between co-regulated subnetworks and contribute to therapy resistance. In a broader sense, OptiCon represents a general framework for systemic identification of synergistic regulators underlying a cellular state transition. OptiCon is implemented using MATLAB.

Update log

The following are changes since the initial release of OptiCon 1.0.

We have optimized the programming of greedy search procedures (including “get_CF.m” and “gen_resultsFun.m”) to improve the OptiCon running efficiency.

I. Input files

OptiCon requires three tab-delimited text files as the input, “DScore.txt”, “GeneExpression.txt”, and “RecurMutant_entrez.txt”.

1) **“DScore.txt”** contains data in two tab-delimited columns. Each row includes a

gene identifier (first column) and its corresponding p-value of differential expression (second column) under two conditions (e.g. diseased vs. healthy). The gene identifier is based on Entrez Gene ID with a prefix "En_". Duplicated gene identifiers are not allowed. Place this file in the "**OptiCon/**" directory.

"OptiCon/InputExample/DScore.txt" shows an example for lung cancer. The p-value was computed using RNA-Seq data generated from tumor tissues and matched normal tissues in 57 lung adenocarcinoma patients. P-value was adjusted for multiple testing using the method of Benjamini and Hochberg. RNA-Seq data were downloaded from the Genomic Data Commons (GDC, https://gdc-portal.nci.nih.gov/).

2) **"GeneExpression.txt"** contains data in a matrix format. Rows represent genes and columns represent samples. Each entry in the matrix contains a gene expression value in a specific sample. Duplicated gene identifiers are not allowed. Place this file into the "**OptiCon/**" directory.

"OptiCon/InputExample/GeneExpression.txt" shows an example of gene expression data in lung cancer. Data source is the same as above.

3) **"RecurMutant_entrez.txt"** contains a list of genes (based on Entrez gene IDs) that are known to harbor recurrent somatic mutations in the cancer type of interest. Duplicated gene identifiers are not allowed. Place this file in the "**OptiCon/output/**" directory.

"OptiCon/InputExample/RecurMutant_entrez.txt" shows an example for lung cancer. The mutation data was obtained from the Catalogue of Somatic Mutations in Cancer (COSMIC) database.

II. Specifying the absolute path of MATLAB program

In the five files below, specify the absolute path of executable MATLAB program in your Linux cluster.

1) "OptiCon/**GreedySearch_input.sh**" (Line 14 and 16).
2) "OptiCon/**gen_m_sh_qsubFiles.py**" (Line 24)
3) "OptiCon/output/**OptiCon_output.sh**" (Line 13)
4) "OptiCon/output/**Gen_SynScoNull.sh**" (Line 3)
5) "OptiCon/output/**gen_m_sh_qsubFiles_randNull.py**" (Line 22)

III. Running OptiCon

1) Copy the "OptiCon/" directory under the "OptiCon_package_v1.1" folder to your working space and set the current working directory to “/Your/path/OptiCon/”. **Run the bash script “GreedySearch_input.sh”.** It is recommended to **use the “qsub”** command on a Linux cluster.

Outputs: MAT-files, MATLAB scripts and bash scripts necessary for Step 2.

Tips: a) Before running this step, please change the value of "e (epsilon)" in the "comp_w.m" file to a Pearson Correlation Coefficient value that corresponds to a two tailed p-value of 0.05 based on the total number of gene expression samples. A useful link for this computation: http://www.danielsoper.com/statcalc/calculator.aspx?id=44

b) Note that one should copy a cleaned “OptiCon/” directory to your working space if a new OptiCon analysis is needed.

2) Set the current working directory to “/Your/path/OptiCon/” (same as Step 1) and **run the bash script “QSUB.sh”** that is generated by Step 1 (**do NOT use "qsub"**).

Outputs: MAT-files necessary for following steps.

Tips: Hundreds of jobs are submitted to your Linux cluster in this step and each job will generate a MAT-file “finTherapTar_relax*.mat”.

3) Set the current working directory to “/Your/path/OptiCon/output/” and **run the bash script “Gen_SynScoNull.sh” (do NOT use "qsub")** to generate a null distribution of synergy scores based on 10 million randomly selected gene pairs from the gene regulatory network.

Outputs: 200 “randSynScore*.mat” files, which collectively constitute a null distribution of synergy scores that will be used to calculate empirical p-values for identified synergistic gene pairs in Step 4.

Tips: 200 jobs are submitted to your HPC cluster in this step and each job will generate a MAT-file “randSynScore*.mat”.

4) Set the current working directory to “/Your/path/OptiCon/output/” (same as Step 3) and **run the bash script “OptiCon_output.sh”**. It is recommended to **use the “qsub”** command on a Linux cluster.

IV. Output

The output file, **“OCN_pairs.txt”**, contains a list of optimal control node pairs (i.e. synergistic key regulators) ranked based on their synergy scores. Original empirical p-values and multiple-testing-corrected p-values are also included.

Appendix

1000 structural control configurations (SCCs) of our constructed gene regulatory network are provided in the "OptiCon/" folder (CF_*.mat). If you want to identify synergistic key regulators using a customized directed network, files in the OptiCon/Gen_SCCs/ can be used to generate a given number of SCCs of your customized network.

1) Format your network data into a tab-delimited file "MyGeneNetwork.txt". Each row represents a directed edge from the node in the first column to the node in the second column.

2) Specify the number of SCCs you want to generate in the Line 7 of the file "gen_diffSCCs.cpp".

3) Pre-define the maximum number of nodes in your customized network in the Line 9 of the file "gen_diffSCCs.cpp".

4) In the Line 17 of the file "OptiCon/Gen_SCCs/gen_SCCs.sh", specify the absolute path of executable MATLAB program in your Linux cluster.

5) Set the current working directory to "Your/path/OptiCon/Gen_SCCs/" and run the bash script "gen_SCCs.sh". It is recommended to use the "qsub" command on a Linux cluster.

6) Run the OptiCon using the four steps described above.

Contact:

Kai Tan, tank1@email.chop.edu
Lin Gao, lgao@mail.xidian.edu.cn
Yuxuan Hu, yuxuan_hu_xd@163.com