**HiFreSP Help**

**Title:** HiFreSP: A novel high-frequency subpathways mining approach to identify robust prognosis gene signatures

**Author:** Chunquan Li

**Email:** [lcqbio@aliyun.com](mailto:lcqbio@aliyun.com)

**Description:** A novel High-Frequency Sub-Pathways mining approach (HiFreSP) to identify robust prognosis gene signatures. The High-Frequency Genes (HFG) and the High-Frequency Pathways (HFP) scores were calculated to mine the prognosis-related sub-pathways, and this method provided the robustness to the noise of the training set and prevent over fit (see details in ‘Materials and Methods’ section of the paper).

**Depends:** R (2.15.2), igraph R package, iSubpathwayMiner R package, survival package

**Details:**

**First step: Executive program: random\_cox\_pathway.R**

**random\_cox\_pathway.R:** construct the bootstrap training sets; identify prognosis-related gene sets and prognosis-related pathway sets; calculate the HFG score and HFP score.

**Input: genedata:** gene expression profile, eg. “genedata\_eg.csv”

**random\_times:** the times of bootstrap processing, eg. “100”

**idconvert:** id convert profile, eg. “convertable.csv”

**Output:** **coxresult**: the result folder of univariable Cox regression analysis in all bootstrap training sets, eg. “coxresult\_1.csv”

**enrichment\_pathway**: the result folder of significant prognosis-related pathway sets in all bootstrap training sets, eg.“signifi\_pathways\_random\_sample1.txt”

**pmRNA**: the folder of genes frequency in the significant prognosis-related pathway, eg. “statis\_path00020.txt”

**pid\_statis:** the frequency of significant prognosis-related pathway, eg, “pid\_statis.txt”

**Second step: Executive program: subpathway\_miner.R**

**subpathway\_miner.R:** based on HFG and HFP scores from random\_cox\_pathway, mine prognosis related sub-pathways

It depends on four functions: annotation.R;

**Dependent function --** annotation.R: get KO sub-pathway annotation

**Input: train\_set:** gene expression profile, eg. “genedata\_eg.csv”

**random\_times:** the times of bootstrap processing, eg. “100”

**T:** the threshold of HFP score, eg. “0.5”

**idconvert:** id convert profile, eg. “convertable.csv”

**robs\_mRNA**: the high-frequency-genes in the significant prognosis-related pathway from the output of random\_cox\_pathway.R, which is “statis\_path00020.txt”

**pid\_names:** the frequency of significant prognosis-related pathway from the output of random\_cox\_pathway.R, which is “pid\_statis.txt”

**Output:**

**subpathway**: the folder of the significant subpathways, eg. “subpathways\_path04530.txt”

In “subpathways\_path04530.txt”, each row of the results represents a subpathway. The meaning of each column is as follows:

**PathwayID--**The identifier of the subpathway

**PathwayName--**The name of the pathway

**annMoleculeRatio--**The ratio of the annotated genes. For example, 9/5909 means that 9 of 5909 genes of interest are annotated to the subpathway

**annBgRatio--**The ratio of the annotated genes of the background. For example, 12/25051 means that 12 genes in 25051 genes of background are annotated to the subpathway

**pvalue--**The P-value of the hypergeometric test

**fdr--**Benjamini-hochberg FDR value

**annMoleculeList--**The molecules annotated to this subpathway

**pdf:** the folder of the significant subpathway’s KM curve in the training set, eg. “train\_path04530\_1.pdf”

**the set of the significant subpathways,** eg. **“**result\_subpathway\_n1s5.txt”

In “result\_subpathway\_n1s5.txt”, each row of the results represents a subpathway. The meaning of each column is as follows:

**pathwayid --**The identifier of the pathway

**subpathwayid--**The identifier of the subpathway

**r\_mRNA--**The genes of interest that are annotated to the subpathway

**ptrain\_KM--**The KM p value of the subpathway in the training set