RAP User Manual

Improving network-based gene prioritization with an integrative strategy

Version 1.0

September 12th , 2016

Authors: Jingjing Zhai†, Yunjia Tang†, Hao Yuan, Longteng Wang, Haoli Shang, Chuang Ma.

Contact:

Dr. Chuang Ma, chuangma2006@gmail.com

Ms. Jingjing Zhai, zhaijingjing603@gmail.com

1 Introduction

Gene prioritization in plants is mainly based on network, while network-based gene prioritization always limits in the number of connection. Here we developed a new method named RAP to prioritize genes, which can be used to perform the gene prioritization in Arabidopsis thaliana 28 non-plant species. RAP improved the predictive accuracy by integrating sequence and network knowledge, using random forest to model, and then aggregating the rank using meta-analysis method. We showcased the utility of RAP by prioritizing flowering-time genes in Arabidopsis and made great success. RAP is written using R programming language, and compiled successfully in Ubuntu V14.04 system.

2 Download

RAP can be used freely for any non-commercial studying or scientific research. It can be available from http://bioinfo.nwafu.edu.cn/software

3 Installation

Supposing that you downloaded the “RAP_1.0.tar.gz” in the directory "/home/zhaijj/download", then open R and install the RAP package with the command:

install.packages(“/home/zhaijj/download/RAP_1.0.tar.gz”, repos = NULL, type="source")

Then you can use command “library(RAP)” to check if it is installed

successfully.

4 Run

Once you have installed the RAP package successfully, then you can run it with the command:

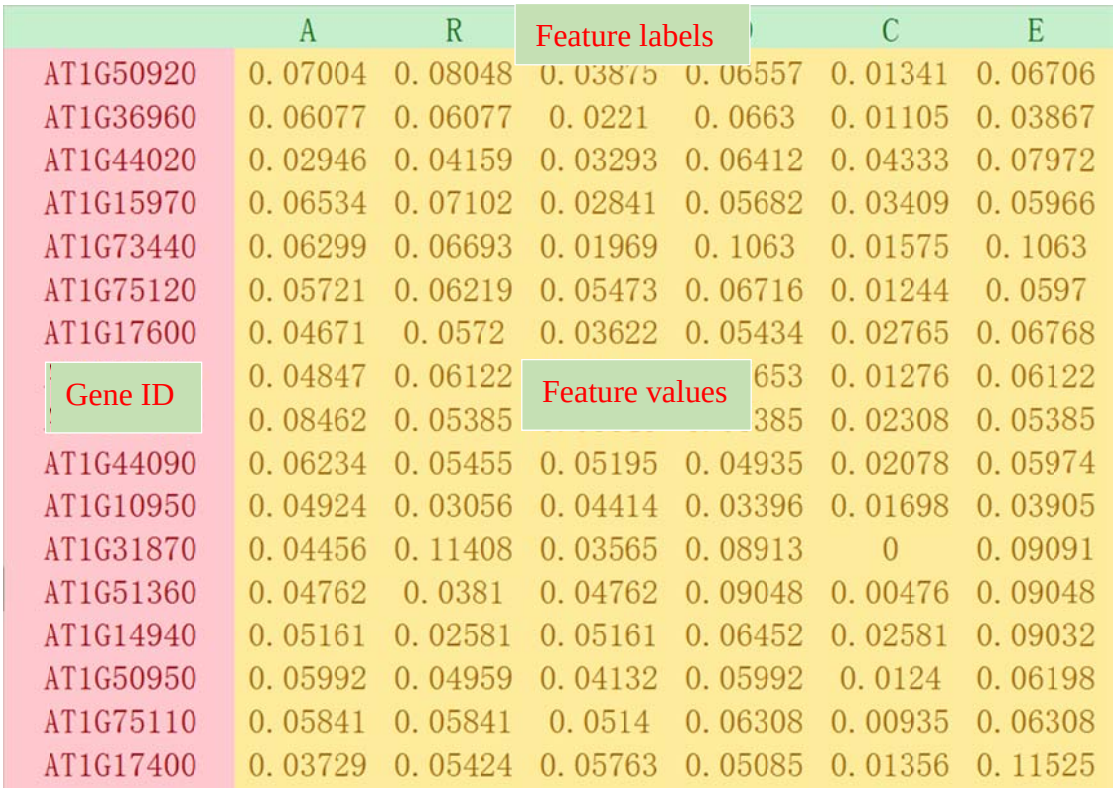
RAP(netPredResult, positives, negatives = NULL, featureSel = TRUE, featureMat = NULL, ProteinSeq, PPIMat, GenomeGeneID, ntree = 500)

The details of each parameter are shown as below:

netPredResult: Full path of gene prioritization results from the network-based gene prioritization algorithms (e.g., AraNet v2), the example is available here:

http://bioinfo.nwafu.edu.cn/wp-content/uploads/AraNetPred.txt

featureMat: A numeric matrix of features(**see Figure 1**) where rows represent genes, cols represent features.



	A	R			C	E
AT1G50920	0. 07004	0. 08048	0. 03875	0. 06557	0. 01341	0. 06706
AT1G36960	0. 06077	0. 06077	0. 0221	0. 0663	0. 01105	0. 03867
AT1G44020	0. 02946	0. 04159	0. 03293	0. 06412	0. 04333	0. 07972
AT1G15970	0. 06534	0. 07102	0. 02841	0. 05682	0. 03409	0. 05966
AT1G73440	0. 06299	0. 06693	0. 01969	0. 1063	0. 01575	0. 1063
AT1G75120	0. 05721	0. 06219	0. 05473	0. 06716	0. 01244	0. 0597
AT1G17600	0. 04671	0. 0572	0. 03622	0. 05434	0. 02765	0. 06768
	0. 04847	0. 06122		653	0. 01276	0. 06122
	0. 08462	0. 05385		385	0. 02308	0. 05385
AT1G44090	0. 06234	0. 05455	0. 05195	0. 04935	0. 02078	0. 05974
AT1G10950	0. 04924	0. 03056	0. 04414	0. 03396	0. 01698	0. 03905
AT1G31870	0. 04456	0. 11408	0. 03565	0. 08913	0	0. 09091
AT1G51360	0. 04762	0. 0381	0. 04762	0. 09048	0. 00476	0. 09048
AT1G14940	0. 05161	0. 02581	0. 05161	0. 06452	0. 02581	0. 09032
AT1G50950	0. 05992	0. 04959	0. 04132	0. 05992	0. 0124	0. 06198
AT1G75110	0. 05841	0. 05841	0. 0514	0. 06308	0. 00935	0. 06308
AT1G17400	0. 03729	0. 05424	0. 05763	0. 05085	0. 01356	0. 11525

Figure 1. Feature matrix shown in Excel

positives: A character vector of positive samples.

negatives: A character vector of negative samples

featureSel: A logical value, where TRUE selecting significant features between positive and negative samples.

ProteinSeq: Protein sequence with a fasta format (**See Figure 2**). Supposing that you have sequences with a fasta format named "protein_sequence.fasta", then you can run this command to read the sequence:

ProteinSeq <- readFASTA("protein_sequence.fasta")

Figure 2. The protein sequence with fasta format

PPIMat: A matrix of PPI (**see Figure 3**), which contains 3 coloums represent protein1, protein2, score respectively, the patameter is not required unless paramter negatives are NULL.

Protein 1	Protein 2	Score
3702. AT1G01010. 1	3702. AT1G01020. 1	964
3702. AT1G01010. 1	3702. AT1G01120. 1	151
3702. AT1G01010. 1	3702. AT1G01380. 1	217
3702. AT1G01010. 1	3702. AT1G02220. 1	298
3702. AT1G01010. 1	3702. AT1G02230. 1	415
3702. AT1G01010. 1	3702. AT1G02250. 1	202
3702. AT1G01010. 1	3702. AT1G03250. 2	193
3702. AT1G01010. 1	3702. AT1G03910. 2	222
3702. AT1G01010. 1	3702. AT1G04780. 1	302
3702. AT1G01010. 1	3702. AT1G06620. 1	153
3702. AT1G01010. 1	3702. AT1G07460. 1	190
3702. AT1G01010. 1	3702. AT1G08050. 1	196
3702. AT1G01010. 1	3702. AT1G08290. 1	220
3702. AT1G01010. 1	3702. AT1G10030. 1	615
3702. AT1G01010. 1	3702. AT1G10220. 1	366
3702. AT1G01010. 1	3702. AT1G10410. 1	195
3702. AT1G01010. 1	3702. AT1G10570. 1	284
3702. AT1G01010. 1	3702. AT1G11220. 1	159
3702. AT1G01010. 1	3702. AT1G12390. 1	230

Figure 3 The protein-protein-interaction dataset

GenomeGeneID: A vector of genome ID, the parameter, the parameter is not required unless parameter negatives is NULL.

ntree: Number of trees to grow when using random forest, the default is 500.

5 Release Note

September 13, 2016, the package of RAP (version 1.0) were released for Windows, Linux and Mac platforms.