

User Guide

for

T1SEstacker

v1.0

01/24/2021

1. Introduction

This manual was prepared for the standalone version of T1SEstacker (version 1.0) and the related modules. For usage of the T1SEstacker webserver, please refer to the HELP page of the T1SEstacker website. The link is: <http://www.szu-bioinf.org/T1SEstacker>.

T1SEstacker could be run as a whole pipeline. Alternatively, the modules could be used independently.

2. Manual of T1SEstacker

1) **System requirement:** Linux , Mac or windows.

2) **Software or package prerequisites:** The following software or packages should be pre-installed and configured into environmental variable path.

Perl 5

R==version 3.3 or later

Python3 version==3.6-3.8

*GO (version 1, if source code compilation needed)

Numpy==1.19.4

biopython==1.78

Pandas version==0.25.3 -1.2.0

Tensorflow == 2.4.0

seaborn==0.9.0

R package: randomForest, e1071,

3) Download and installation of T1SEstacker:

The T1SEstacker package (T1SEstacker.v1.0) for Linux or Mac or windows system could be downloaded from the website: <http://www.szu-bioinf.org/tools/T1SEstacker>.

Decompress the “T1SEstacker.tar.gz” and get into ~/T1SEstacker/ from terminal. If the pre-compiled version does not work, try to re-compile all the GO scripts in the “codes” ,and replace the binary files in the “bin” sub-folder of each module with the newly compiled ones, respectively. Compiling GO scripts:

```
$ cd ~/T1SEstacker/MODULE_NAME/bin
```

```
$ go build ../codes/xxx.go
```

“xxx.go” is the script in “codes” folder. Compile all the scripts.

4) Input files:

There is a **necessary protein sequence file** .The ‘protein sequence file’ *T1SEstacker* requires them to be FASTA-formatted, as exemplified by the demonstrated “test.fasta” (Fig 1).

```
1 >CAA48711.1 apxIIIA [Actinobacillus pleuropneumoniae]
2 MSTWSSMLADLKKRAEEAKRQAKKGYDVTKNGLQYGVSQAKLQALAAGKAVQKYGNKLVVIPKEYDGSV
3 GNGFFDLVKAEEELGIQVKYVNRNELEVAHKSLGTADQFLGLTERGLTFAPQLDQFLQKHSKISNVVGS
4 STGDVSKLAKSQTIISSGIQSVLGTVLGINLNEAIISSGSELELAEAGVSLASELVSNIAKGTITIDAF
5 TTQIQNFGLVENAKLGGVGRQLQNISSGALSKTGLGLDIISSLLSGVTASFALANKNASTSTKVAAGF
6 ELSNQVIGGITKAVSSYILAQRLAAGLSTTGPAALIASISLAISPLAFLRVADNFRSKEIGFAERF
7 KKLGYDGDKLLSEFYHEAGTIDASITTTISTALSAIAAGTAAASAGALVGAPITLLVTGITGLISGILEFS
8 KQPMLDHVASKIGMKIDWEKKYGNFYFENGYDARHKAFLEDSFSLSSFNKQYETERAVLITQQRWDEY
9 IGELAGITGKGDKLSGKAYVDYFQEGKLEKKPDDFSKVVFDPTKGEIDISNSQSTLLKFVTPLLTPG
10 TESRERTQTKYEYITKLWVGKDKWVWVGKDKGAVDYDYNLIQHAHISSSVARGEYREVRLVSHLGN
11 GNDKVFLAAGSAETHAGEGHDVYYDKDTGTLVIDGTKATEQGRYSVTRELSGATKILREVIKNQKSAV
12 GKREETLEYRDYELTQSGNSNLKAHDELHSVEEITGSMQRDEFKGSKFRDIFHGADGDDLINGNDGDDIL
13 YGDKGNDLREGDNGDQLYGGEGNDKLLGGNGNNYLSGGDGNDELQVLGNGFNVLRGKGDDKLYGSSGS
14 DLLDGGEGNDYLEGGDGSDFYVYRSTSGNHTIYDQKSSDLKLYLSDFSFDRLLEKVDNLLVRSNES
15 SHNNGVLTIKDWFKEGNKYNHKIEQIVDKNGRKLTAENLGTYFKNAPKADNLLNYATKEDQNESLSSLK
16 TELSIIITNAGNFVAKQGNITGAALNNEVNKIISSANTFATSQLGSGMGLPSTNVNSMMLGNLAR
17 AA
18
19 >CAA45858.1 alkaline protease [Pseudomonas aeruginosa PA01]
20 MSSNSLALKGRSDAYTQVDNFLHAYARGDELVNGHPSYTVDAQAEQILREQASWQKAPGDSVLTLSYSF
21 LTKPNDFFNTPKYVSDIYSLGKFSAFSAQQQAQKLSLQSWSDVTNIHFVADAGQDQDGLTFGNFSSSV
22 GGAFAFALPDPVDPALKGQSWYLINSSYSANVNPANGNYGRQTLTHEIGHTLGLSHPGDYNAGEGDPYAD
23 ATYAEOTRAYSVMYSYWEQNTGQDFKAYSSAPLLDDIAAIQKLYGANLTTRTGDTVYGFNSNTERDFYS
24 ATSSSSKLVFVMDAGGNDLDFSGFSQNKINLNEKALSDVGGKGNVIAAGVTVENAIGGSGSDLLI
25 GNDVANVLKGGAGNDILYGGGLGADQLWGGAGADTFVYGDIAESSAAAPDTRDFVSGQDKIDLGLDAFV
26 NGGLVLQYVDAFAGKAGQAILSYDAASKAGSLAIDFSGDAHADFAINLIGQATQADIVV
27
28 >CAA35178.1 unnamed protein product [Rhizobium leguminosarum]
29 MNIKGSNMGSIKGSPENDIIDGGKKNWIDAGNGDDRIKAGDGDQDITAGPGHDIVWAGKGSVDIHADG
30 GDDLLYSDASYPLVYDTPHRVIPHSSEGDDVLYAGPGSDILVAGDGADVLTGDDGDADFVRFHDPVMGT
```

Fig 1. Input sequence file format - FASTA

5) Running T1SEstacker in one command line:

get into ~/T1SEstacker/ through command line:

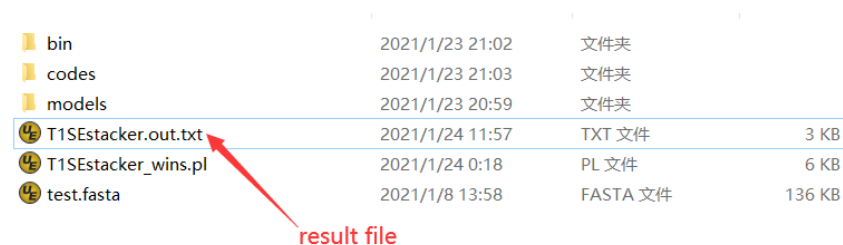
```
$ cd ~/T1SEstacker/
```

running T1SEstacker through command line:

```
$ perl T1SEstacker.pl test.fasta
```

6) Output file format:

The prediction results are given in ~/T1SEstacker. 'T1SEstacker.out.txt' gives the prediction results. (Fig 3).



bin	2021/1/23 21:02	文件夹	
codes	2021/1/23 21:03	文件夹	
models	2021/1/23 20:59	文件夹	
T1SEstacker.out.txt	2021/1/24 11:57	TXT 文件	3 KB
T1SEstacker_wins.pl	2021/1/24 0:18	PL 文件	6 KB
test.fasta	2021/1/8 13:58	FASTA 文件	136 KB

Fig 3. Prediction result files and their details

'T1SEstacker.out.txt' contains 3 columns: protein ID, the total score of five models, and the classification with default cutoff of each model and the classification according to the customized or default (0.6) cutoff. (Fig 4).

prot	voting	pred
CAA48711.1	1	T1SE
CAA45858.1	1	T1SE
CAA35178.1	1	T1SE
CAA39137.1	1	T1SE
sp P22522.1 CEAV_ECOLX	0	nonT1SE
sp P08715.1 HLYAP_ECOLX	1	T1SE
sp P22542.1 HSTI_ECOLX	0	nonT1SE
AAA63637.1	1	T1SE
AAA24860.1	1	T1SE
AAA25881.1	1	T1SE
AAA25882.1	1	T1SE
BAA02519.1	0.8	T1SE
AAA99902.1	1	T1SE
CAA50501.1	1	T1SE
CAA49611.1	1	T1SE
AAA16444.1	0.2	nonT1SE
AAA81002.1	1	T1SE
sp Q07162.1 PRTG_DICCH	1	T1SE
CAB01938.1	1	T1SE
AAB64093.1	1	T1SE
CAA05794.1	1	T1SE
AAC33448.1	1	T1SE

Fig 4. The final prediction result of T1SEstacker