**Supplementary Method S3.** Manual of the CNN-T4SE

**TITLE:** Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery

**DESCRIPTION:** The CNN-T4SE is a tool that can be run in *windows* and *linux* operating systems. It is used to identify effector proteins through the amino acid sequence, evolutionary information in the form of a position-specific scoring matrix, secondary structure and solvent accessibility data of the proteins. After prediction, this tool provides CSV files that contains the protein IDs, prediction probabilities and prediction results.

**OPERATION PROCEDURE**

**1. Download the CNN-T4SE**

The CNN-T4SE tool is provided in the website of https://idrblab.org/cnnt4se/. To use it, firstly, download the CNNT4SE.zip file for *windows* user, or CNNT4SE.tar.gz file for *linux* user, then extract the file to any directory in your computer.

**2. Prepare the Input Files**

The input files of CNN-T4SE includes amino acid sequence file, the secondary structure sequence file, the solvent accessibility sequence file and the PSSM files. Both of the secondary structure sequence file and the solvent accessibility sequence file can be generated by a protein structure prediction tool SCRATCH (a local version, *SCRATCH-1D_1.2.tar.gz*, were provided in the website of https://idrblab.org/cnnt4se/). The PSSM files can be generated by an online tool POSSUM, which is provided in the website of http://possum.erc.monash.edu/. After getting the input files, put them into the CNNT4SE directory, the PSSM files should be put into the *pssm_files* subdirectory.

**2.1 Installing and using SCRATCH**

2.1.1 To use SCRATCH, you should firstly to download *SCRATCH-1D_1.2.tar.gz* file and extract the it to any directory in your computer.

*tar -zxvf SCRATCH-1D_1.2.tar.gz*

2.1.2 Then go into the *SCRATCH-1D_1.2* directory through the command:

*cd SCRATCH-1D_1.2*

2.1.3 Before installing the software, you should ensure that you have *gcc* and *perl* installed on your system. Then you can install it by the command:

*perl install.pl*

2.1.4 After that you can get an executable file *run_SCRATCH-1D_predictors.sh* in the bin directory. To generate the secondary structure sequence file and the solvent accessibility sequence file, you should firstly go to the directory of the protein sequence (fasta format) that you want to use to generate these two files. The execute the command

*/your/path/to/run_SCRATCH-1D_predictors.sh input.fasta name.out 4*

2.1.5 After that 4 output files can be get,

- name.out.ss

- name.out.ss8

- name.out.acc

- name.out.acc20

The name.out.ss and name.out.acc are the secondary structure sequence file and the solvent accessibility sequence file which can be used to predict Type IV effectors.

2.1.6 The most common issue encountered by users of the previous SCRATCH-1D releases comes from the incompatibility between the blast binaries delivered with SCRATCH-1D and the operating system used to run the software. The 32 bit linux version of blast provided by default in the *pkg* directory are not compatible with any type of operating system so in many cases these blast binaries need to be replaced by the ones specifically compiled for the given operating system. The blast is put in the *pkg* directory, you replace it with a satisfied version of blast, it can be download at *ftp://ftp.ncbi.nlm.nih.gov/blast/executables/legacy.NOTSUPPORTED/2.2.26/*

2.1.7 There is README.TXT file that containing the user manual of SCRATCH in the *SCRATCH-1D_1.2* directory.

**2.2 Using the POSSUM to generate PSSM files**

2.2.1 Go to the website of http://possum.erc.monash.edu/, click the *Go to use it* button. Then copy your protein sequences (fasta format) to the *Examples* panel and input your e-mail in the *E-mail* panel. After that, you can click the *submit* button to submit your job with all the parameters set as the default. And you will get a job id and a job name. You should submit no more than 500 sequences at once, and the length of the sequence should more than 50 and less

than 5000.

2.2.2 When the job is finished, an e-mail will be sent to you. Click the url in the e-mail and go to the job result page. You can also find your job in the job list page (http://possum.erc.monash.edu/jobList.jsp) according to the job name or job id. Here, you can see the status of your jobs and go to the job result page by clicking the *Click* in the *Detail* columns after the job is completed. In the job result page, you can download the PSSM files in a zip package. Each protein will generate a PSSM file, put them into the *pssm_files* directory in CNNT4SE file, then you can use it to predict effectors with our CNNT4SE software.

**3. Predict Effector proteins**

For windows users, firstly, open cmd and change to the CNNT4SE directory, then execute the following command to do the prediction:

*predict.exe mode input_files*

For linux users, just change to the CNNT4SE directory and execute the following command:

*./predict mode input_files*

There are 4 values that can be taken by the *mode* parameter, that is *PSSSA, PSSM, Onehot and Vote*. The *input_files* is determined by the *mode* you choose.

3.1 If you want to predict the effector by the secondary structure and the solvent accessibility, the *mode* should be the *PSSSA* and the *input_files* should be the secondary structure sequence file (*input.ss*) and the solvent accessibility sequence files (*input.acc*).

*predict.exe PSSSA input.ss input.acc (windows)*

*./predict PSSSA input.ss input.acc (linux)*

3.2 If you want to predict by PSSM of proteins, the *mode* should be the *PSSM* and the *input_files* parameter does not need to be specified, just put your PSSM files in the *pssm_files* subdirectory.

*predict.exe PSSM (windows)*

*./predict PSSM (linux)*

3.3 If you want to predict by the sequence feature denoted by a one-hot encoding methods, the *mode* should be the *Onehot* and the *input_files* parameter should be the protein sequence file (as the same format as *Sample.fasta*).

*predict.exe Onehot input.fasta (windows)*

*./predict Onehot input.fasta (linux)*

3.4 If you want to predict by all the three kinds of protein features, the *mode* should be the *Vote* and all feature profile should be specified. You should use the protein ids that are the same as the protein ids in the *input.fasta* file to rename the PSSM files. The order of protein sequences and the protein ids in *input.ss input.acc* and *input.fasta* should be consistent.

*predict.exe Vote input.ss input.acc input.fasta (windows)*

*./predict Vote input.ss input.acc input.fasta (linux)*

For example, to predict the effectors in the sample files by all the three kinds of protein features, the following commad should be executed:

*predict.exe Vote Sample.ss Sample.acc Sample.fasta (for windows)*

*./predict Vote Sample.ss Sample.acc Sample.fasta (for linux)*

**4. The Results of Prediction**

After executing the prediction command, a *Result.csv* file will generate in the CNNT4SE directory. The file contains three columns values, the first column is the protein IDs, the second is the prediction probabilities and the third is the prediction results of each protein. The protein with prediction probability > 0.5 is regarded as the effector protein.