User Guide

Table of Contents

Summary 2
Installation 3
Required 3
GCC installation 3
Wget installation 3
R and perl 3
Pfam installation 4
Hmmer3.0rc2 installation 4
CPAN package installation 4
BioPerl installation 4
Implementation of Pfam 5
LINUX and Mac: 5
Emboss suite installation 5
blastp installation 5
Principle and operation 6
Homology and Pfam Databases 8

Summary

Type IV Effectors (T4Es) are proteins produced by pathogenic bacteria to manipulate host cell gene expression and processes, divert the cell machinery for their own profit, and circumvent the immune responses. T4Es have been characterized for some bacteria but many remain to be discovered. To help biologists to identify putative T4Es from the complete genome of α- and γ-proteobacteria, we developed a PERL-based command line bioinformatics tool called S4TE. The tool predicts and ranks T4Es candidates by using a combination of 13 sequence characteristics, including homology to known effectors, homology to eukaryotic domains, presence of subcellular localization signals or secretion signals, etc. S4TE software is modular, and specific motif searches are run independently before combining the outputs to generate a score and sort the strongest T4Es candidates. The user can adjust various search parameters such as the weight of each module, the selection threshold or the input databases. The algorithm also provides a GC% and local gene density analysis, which strengthens the selection of T4Es candidates. S4TE is a unique predicting tool for type IV effectors, with utility upstream from experimental biology.

Installation

The system was tested on Ubuntu 11.10 and Mac OS 10.7 (Lion).

Required

Some tools are not present in the UNIX environment and are required for good installation. All facilities described below are from UNIX terminal.

GCC installation

GCC is a C compiler; it will install some useful programs to operate S4TE

GCC pkg download:
```
curl http://cloud.github.com/downloads/kennethreitz/osx-gcc-installer/GCC-10.7-v2.pkg > GCC-10.7-v2.pkg
```
pkg installation:
```
sudo /usr/sbin/installer -verboseR -lang fr -pkg ./GCC-10.7-v2.pkg -target /Volumes/Macintosh\ HD
```

Command line `sudo /usr/sbin/installer` installs GCC. This step requires the superuser password of the user.

Wget installation

wget can download ftp files. It has the same function as curl but is more efficient.

Package download:
```
curl ftp://ftp.gnu.org/gnu/wget/wget-1.9.tar.gz > wget.tar.gz
```
untar folder:
```
tar zxvf wget.tar.gz
```
installation:
```
cd wget/
./configure
make
sudo make install
```

R and perl

These two programming languages are essential to properly operate the program.

S4TE requires Rscript, present in recent versions of R, for run R scripts. If this program is not present, reinstall a newer version of R.

To test whether Rscript is installed on the computer, you must run `Rscript`. If help is displayed, RSCRIPT is installed; otherwise, you need to install a recent version of R.

Pfam installation

The Pfam software requires several programs to be functional.

Hmmer3.0rc2 installation

LINUX:

```
wget ftp://selab.janelia.org/pub/software/hmmer3/3.0rc2/hmmer-3.0rc2-linux-intel-ia32.tar.gz
tar zxvf hmmer*
cd hmmer*/
./configure
make
make check
sudo make install
```

Mac:

```
wget ftp://selab.janelia.org/pub/software/hmmer3/3.0rc2/hmmer-3.0rc2-macosx-intel.tar.gz
tar zxvf hmmer*
cd hmmer*/
./configure
make
make check
sudo make install
```

CPAN package installation

LINUX and Mac:

First, the user has to start CPAN (Comprehensive Perl Archive Network) with:

```
sudo CPAN
```

And run Moose and IPC::Run installation with command lines:

```
install Moose
install IPC::Run
```

BioPerl installation

LINUX and Mac:

BioPerl is a package compiling a large number of bioinformatics tools. It is useful for Pfam and for some modules of S4TE program.

Package download and installation:

```
wget http://bioperl.org/DIST/BioPerl-1.6.1.tar.gz
tar xvfz BioPerl-1.6.1.tar.gz
cd BioPerl-1.6.1/
perl Build.PL
```

```
./Build test
sudo ./Build install
```

Implementation of Pfam

LINUX and Mac:

PfamScan.tar.gz download:

```
wget ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/PfamScan.tar.gz
tar xzvf PfamScan.tar.gz
cd PfamScan/
```

Then just move the executable PfamScan.pl in /usr/bin/ folder of your computer with:

```
sudo cp ./pfam_scan.pl /usr/bin/
```

and Bio/ folder in a current directory of Perl for example /usr/local/share/perl/<current_version_(5.12.4)>/
to know the current version of Perl: perl --version

```
sudo cp -R Bio/ /usr/local/share/perl/5.12.4/
```

The Pfam program is now installed and used with the command line `pfam_scan.pl`.

Emboss suite installation

Package download:

```
wget ftp://emboss.open-bio.org/pub/EMBOSS/EMBOSS-6.5.7.tar.gz
```

Installation:

```
tar xzfv EMBOSS-6.5.7.tar.gz
cd ./ EMBOSS-6.5.7.tar.gz
./configure –without -x
make
make check
sudo make install
```

blastp installation

Package download:

LINUX:

```
wget ftp://ftp.ncbi.nih.gov/blast/executables/blast+/2.2.26/ncbi-blast-2.2.26+-ia32-linux.tar.gz
```

Mac:

```
wget ftp://ftp.ncbi.nih.gov/blast/executables/blast+/2.2.26/ncbi-blast-2.2.26+-universal-macosx.tar.gz
```

Installation:

```
tar xzfv ncbi-blast-2.2.26+-*.tar.gz
cd ncbi-blast-2.2.26+-ia32-linux/
```

```
./configure
make
make check
sudo make install
```

Principle and operation

S4TE is an automatic research pipeline for Type VI secretion system effectors. It works on all γ- and α-proteobacteria genomes. For use, the database of the genome <Name_of_the_genome_folder> must be 4 distinct files:

- Genome.nucl:

  Genome sequence in FASTA Format

- Genome.an:

  A csv file (with ; as separator) containing the gene ID; the position of the first nucleotide of coding sequence; the position of the last nucleotide; the sense or antisense status; before use in S4TE, this file needs to be sorted in ascending order from the first nucleotide position.

- Genome.prot:

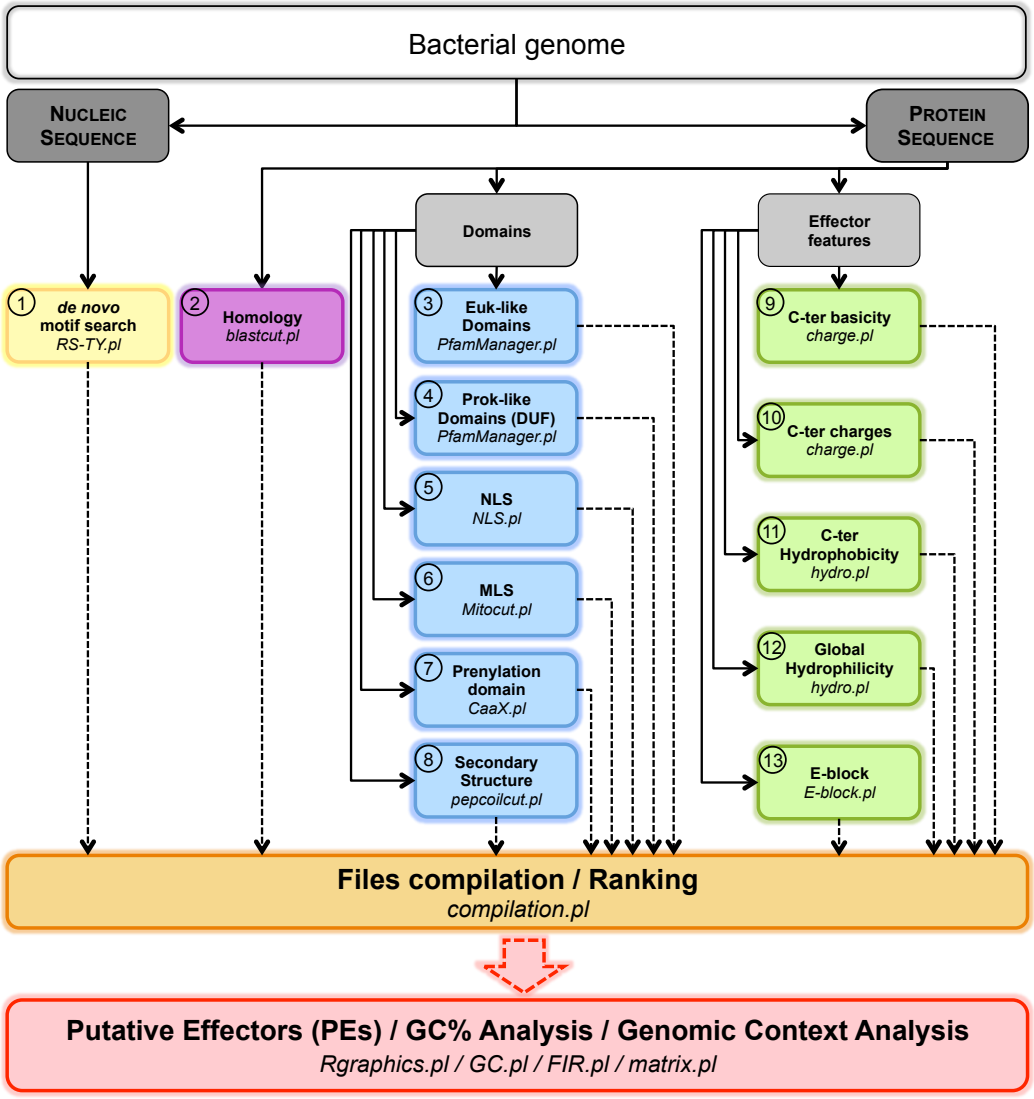
  Fasta file containing all proteins encoded by the genes of the bacterium

- Genome.csv:

  File built from Genome.prot file using the tool nomprot.pl present in the ~ / S4TE/DataBases/Genome/Tools folder

  With the command line `./nomprot.pl -g < Name_of_the_genome_folder >`

Once the database genome is created, simply run the program S4TE.pl to search effectors. The program consists of 10 independent modules to explore the 13 characteristics of putative effectors. It consists of a module for the consensus motif RSTY in the promoters, three modules for search of the five characteristics of secretion (basicity C-ter, charges C-ter, hydrophobicity C-ter, overall hydrophilicity, E-block), five modules for the six domains search (eukaryotic domain, DUF domain, NLS, MLS, prenylation domain, coiled-coils domain) and a homology search module as shown in the chart below.



Each module creates a .txt file in a folder
way_to_S4TE/S4TE/Jobs/job<Name_of_genome_folder><year><month><day><hour><minutes>

All these results are compiled in compilationFile.txt and result.txt files in the same folder.

To start S4TE, go to the S4TE folder cd way_to_S4TE/S4TE/ and launch the following command line ./S4TE.pl -g <Name_of_genome_folder>; this command line launches S4TE automatically with the default options. Options are available to launch S4TE to change search parameters.

The -c option allows you to delete a module on S4TE pipeline search. The threshold and weighting will be automatically recalculated.

The -w option allows you to impose a new weighting code (13-digit code corresponding to the weighting of each module in the order of execution of the program as shown in the chart from left to right). The threshold will be recalculated using the new weighting

The -t option imposes a threshold of acceptance on the program.

Homology and Pfam Databases

The databases used by S4TE are homology and Pfam databases.

The basis of homology is created from a file of protein sequence effectors (effector_db.txt) formatted with the command line:

```
makeblastdb -in effector_db.txt -dbtype prot
```

The Pfam database must be constructed from the original ftp downloadable database:

```
wget ftp://ftp.sanger.ac.uk/pub/databases/Pfam/releases/Pfam26.0/Pfam-A.hmm*.gz
```

This command line will download Pfam-A.hmm.gz and Pfam-A.hmm.dat.gz files. Pfam-A.hmm.gz is very heavy and cannot be opened and manipulated easily. The file Pfam-A.hmm.dat.gz references all references of Pfam motifs. It is much lighter and easier to handle. The tool hmmcut.pl was written to reformat Pfam-A.hmm from Pfam-A.hmm.dat modified file (containing only the desired patterns searched with S4TE pipeline). In the code, modify the pathways to access the different files (see help). Then, replace Pfam-A.hmm by the newly created Pfam-A.hmm.new, and rename it Pfam-A.hmm.

The database Pfam-A.hmm will then be formatted to make it usable. To do this, run the command line:

```
./hmmpress Pfam-A.hmm
```