Search\_TPS - A Quick Guide and Tutorial

1. **Introduction**

Search\_TPS is a Perl script that uses PFAM models [Mistry *et al.*, 20201] and class-specific profile HMMs to detect and classify TPS sequences in a set of input sequences (Fig. 1). First, search\_TPS performs similarity searches using *hmmsearch* program (HMMER package [Eddy *et al*., 2011]) of PFAM models against the input sequences to identify possible TPS sequences. From the results, the program selects sequences presenting similarity with at least two PFAM models. Using these sequences, the program then performs similarity searches using hmmsearch with class-specific profile HMMs. Sequences presenting similarity with at least two profile HMMs of a same class are selected. These sequences can be classified as selected with high confidence if they present a score equal to or greater than the cutoff score of the HMMs profile, or as selected with low confidence, if they have a score lower than the cutoff scores of the HMMs profile and greater than a low confidence cutoff value (default = 100). Based on the scores obtained by selected sequences, search\_TPS classifies them in monoTPS, diTPS and sesquiTPS.

![Diagrama

Descrição gerada automaticamente]()

**Figure 1.** Search\_TPS workflow. (A) As a first step, search\_TPS performs similarity searches using the hmmsearch program of PFAM models against the input strings. Sequences that show similarity with at least two models are selected for the next step. (B) The selected sequences are then used as input for similarity searches with hmmsearch using the class-specific profile HMMs. Sequences that present similarity with at least two specific profile HMMs for the same class are selected. (C) Finally, the selected sequences are classified based on the highest score value obtained in the similarity searches.

1. System requirements

search\_TPS is a program developed in Perl language. It can be used in a POSIX compliant operating system (for example: UNIX and Linux distributions) with an installed Perl interpreter (<http://www.perl.org)>.

1. Third-part programs

Search\_TPS only requires the program hmmsearch (HMMER3 package - <http://hmmer.org/)>. This program is used to run similarity searches using PFAM models and class-specific profile HMMs against a set of sequences. This program must be located in a directory listed in the PATH of the operating system.

1. How to cite

If you will use search\_TPS program for your publication, please cite it as:

* search\_TPS program (developed by Liliane S. Oliveira, UTFPR, Brazil, unpublished).

1. search\_TPS parameters:
   1. Mandatory parameters:

Search\_TPS has five mandatory parameters:

-d <class-scpecific\_phmms\_dir> : Directory containing specific

profile HMMs for monoTP, diTP and

sesquiTP.

-i <input\_fasta\_file> : Fasta file containing the sequences

to be searched.

-p <pfams\_dir> : Directory containing the PFAMs

models.

-s <search\_options> : Group of TPS to be searched:

1 - all (monoTP, diTP and sesquiTP)

2 - monoTP

3 - diTP

4 - sesquiTP

-t <table\_file> : File containing the cutoff scores to

be used to select the results of

each HMM profile.

* 1. Optional parameters:

-cpu <num> : Number of threads to be used by

hmmsearch.

-h|help : Show this help message.

-l <low\_confidence\_score> : Low confidence score to be considered

(default = 100).

-o <dir> : Output directory (default = output\_dir).

-v|version : Version.

1. Running search\_TPS

In this tutorial we will execute the program search\_TPS to detect TPS sequences for monoTPS, diTPS and sesquiTPS classes in a set of Arabidopsis thaliana proteins. In this execution, we will use three TPS models from PFAM database (PF01397, PF03936 and PF19086) and 27, 24 and 60 profile HMMs specific for monoTPS, diTPS and sesquiTPS classes respectively. We will also use a table containing cutoff scores defined for each class-specific model. To perform the program, we should type the following command line:

search\_TPS.pl -i Phytozome\_Atha\_Araport11.fa -d class\_specific\_dir -p PFAMs\_dir -t scores\_table\_dir/all\_scores.csv -o output\_dir -s 1

Files and directories used in this execution:

* Phytozome\_Atha\_Araport11.fa: file containing Arabidopsis thaliana proteins obtained from Phytozome database (<https://phytozome.jgi.doe.gov/pz/portal.html>).
* PFAMs\_dir: directory that contains the PFAMs models PF01397, PF03936 and PF19086.
* class\_specific\_dir: directory containing class-specific profile HMMs for monoTPS, diTPS and sesquiTPS classes.
* scores\_table\_dir/all\_scores.csv: CSV file containing the cutoff scores for all specific profiles HMMs. These values were manually calculated based on the results of similarity searches using hmmsearch program of the models against the training set (sequences used to build the models).

1. Output files and directory

search\_TPS program execution generates several files and subdirectories in an output directory:

* file.log: file that reports the steps of the execution.
* error.log: file containing the error messages. If the execution ends without errors, this file remains empty.
* PFAM\_dir: this directory contains files generated by hmmsearch execution with the PFAM models, a tabular file listing the header of the selected sequences and a FASTA file containing the selected sequences.
* The subdirectories monoTP\_dir, diTP\_dir and sesquiTP\_dir contains the following files:
  + \*hmmsearch.txt: This file contains all results found by the execution of hmmsearch program for each profile HMM.
  + \*hmmsearch\_high.tab: This file contains the results found by the execution of hmmsearch program for each profile HMM selected by the cutoff score of the model (high confidence results).
  + \*hmmsearch\_low.tab: This file contains the results found by the execution of hmmsearch program for each profile HMM selected by the low confidence cutoff score (low confidence results).
  + high\_confidence\_results.csv: This file contains all sequences that presented similarity with at least of two models of the same class with high confidence. In this case, a sequence could be classified in more than one class with high and low confidence.
  + low\_confidence\_results.csv: This file contains all sequences that presented similarity with at least of two models of the same class with low confidence. In this case, a sequence could be classified in more than one class with high and low confidence.
  + high\_confidence\_final\_results.csv: This file contains the sequences that presented similarity with at least of two models of the same class with high confidence without redundance. In this case, a sequence is classified in only one class, and it is not present in the final file of sequences classified with low confidence.
  + low\_confidence\_final\_results.csv: This file contains the sequences that presented similarity with at least of two models of the same class with low confidence without redundance. In this case, a sequence is classified in only one class, and it is not present in the final file of sequences classified with high confidence.

All files described in this document are available in github (<https://github.com/liliane-sntn/TPS>). There, you will find FASTA sequences and TPS results found by search\_TPS program for Coffea arabica (Cara files), Coffea canephora (Ccan files), Coffea eugenioides (Ceug files) and Quillaja Saponaria (Qsap files).

1. References

Eddy, S.R. (2011) Accelerated Profile HMM Searches. PLoS computational biology, 7, e1002195.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L., ... & Bateman, A. (2021). Pfam: The protein families database in 2021. Nucleic Acids Research, 49(D1), D412-D419.