e-Finder

A tool to find multigene elements in assembled sequences using profile HMMs

A Quick Guide

University of São Paulo

e-Finder - A Quick Guide and Tutorial

1 Introduction

`e-Finder` is a generic tool for detection and extraction of multigene elements from assembled genomes using profile HMMs. `e-Finder` executes `hmmsearch` program (HMMER package) to run similarity searches using profile as queries against translated sequences of the assembled genomes. Any region containing a cluster of genes can be detected, such as transposons, CRISPR-Cas systems, prophages, operons, etc.

2 Main features

- Fully configurable
- FASTA files containing finished or partially assembled genomes can be used as input
- Previously run `hmmsearch` results can be reprocessed with new sets of parameters
- Score and e-value cutoff values can be used to set the stringency of the search

3 Before using `e-Finder`

3.1 System requirements

`e-Finder` was developed in Perl language and can be used in any POSIX compliant operating system such as UNIX and Linux distributions with an installed Perl interpreter (http://www.perl.org).

3.2 Third-party programs

`e-Finder` requires the following programs and databases:

- `transeq` (EMBOSS package - http://emboss.sourceforge.net/). This program is used to translate assembled nucleotide sequences into the six possible frames.

- `hmmsearch` (HMMER3 package - http://hmmer.org/; Eddy, 2011). This program is used to run similarity searches of profile HMMs against metagenomic datasets. The program must be located in a directory listed in the PATH of the operating system.
- `tblastn` – (BLAST package – http://blast.ncbi/nlm/nih/gov); Altschul *et al*., 1997). The program is used to detect the coding regions corresponding to the positive HMMs in the assembled sequences.

3 Understanding `e-Finder`'s workflow

3.1 Input data

`e-Finder` uses two types of input data: a set of profile HMMs and one or more FASTA files containing assembled sequences (Figure 1A). In addition, a tabular file containing a list of dataset identifiers and the respective organism names can optionally be used. This information, if provided, is used to generate a final CSV report.

3.1.1 profile HMMs

Profile HMMs are used to identify the corresponding genes in the assembled sequences. Single or multiple protein markers can be used to detect the elements, each one represented by one or more profile HMMs. For instance, casposons are mobile genetic elements composed of a variable number of genes (Krupovic *et al*., 2014) but always presenting *cas1* (Cas1 endonuclease) and *polB* (PolB-like RNA polymerase) genes. Thus, one can use profile HMMs constructed from proteins coded by these genes to fish casposon elements in bacterial and archaeal genomes. Also, if multiple profile HMMs are available for each protein, they can all be used in combination. This strategy may increase the sensitivity of detection for more remote elements, as different protein domains may present variable rates of divergence and very often such rates are not known *a priori*. `e-Finder` can identify cutoff score tags inserted in the profile HMMs and use their values in the sequence selection step (see below for more details).

3.1.2 Assembled sequence datasets

As input, `e-Finder` can use FASTA files containing sequences derived from either single or multiple genomes. Alternatively, `e-Finder` can also use a directory containing multiple subdirectories, each one specific for a given genome, which in turn contain the corresponding FASTA files. This input option is particularly convenient when using genomic data from PATRIC (https://www.patricbrc.org/ – Wattam *et al*., 2017), the Bacterial Bioinformatics Resource Center, a comprehensive repository of assembled genomes that employs this data storage structure. Other genomic and metagenomic datasets can also be used, provided that they are previously assembled.

3.2 Detection of candidate sequences

`e-Finder` uses profile HMMs as query in similarity searches against assembled sequences (Figure 1B). Since `hmmsearch` does not automatically translate nucleotide sequences (like `tblastn` does), `transeq` program is invoked to translate all sequences into the six possible open reading frames. Once this step is performed, `e-Finder` then executes `hmmsearch` to run the searches with its default parameter `-S 10`, which corresponds to an e-value of 10, a low-stringency condition. While this choice ensures a high sensitivity of detection, the user can still define a second score or e-value to be used as threshold for the sequence selection step. Different secondary thresholds can be used in multiple `e-Finder` executions to arbitrarily increase specificity, with no need to rerun the `hmmsearch` search. In fact, if `e-Finder` detects a previous similarity search file, it will only parse this file out to retrieve those results that comply to the newly established threshold. This feature allows to skip the slow similarity search execution, saving precious time. Alternatively, threshold values can also be specifically hardcoded in each profile HMMs as cutoff score tags (e.g. `CUTOFF SCORE 45.3`). `e-Finder` automatically identifies this tag in all models, running `hmmsearch` with parameter `-T` (`45.3` in this example), thus allowing custom cutoffs to be used for each corresponding model. The `CUTOFF SCORE` is a proprietary tag that has no effect on `hmmsearch` program itself but allows `e-Finder` to invoke it with the most appropriate threshold for each profile HMM.

3.3 Selection of positive sequences

Once the positive sequences have been detected, `e-Finder` then checks if user-defined criteria for syntenic context have been fulfilled (Figure 1C). Thus, each sequence must contain a minimum number of markers and the appropriate intergenic distances, including or not the possibility of overlapping genes. Beside gene composition, the user can also specify whether or not a strict gene order must be found. Regions containing the multiple gene markers are extracted together with size-defined 5'and 3' flanking regions and submitted to size filtering. By using a selected genetic code, ORFs coding for the protein sequences detected by the models are then identified, extracted and conceptually translated into full-length protein sequences.

3.4 Sequence extraction and report generation

In the final phase (D), `e-Finder` stores the nucleotide and protein sequences of each selected region. Also, the program generates and stores a CSV spreadsheet file listing all elements found in each assembled sequence and their corresponding markers, coordinates and intergenic distances, among other features.

4 Understanding `e-Finder` parameters

4.1 Mandatory parameters:

`e-Finder` has two mandatory parameters that must be specified by the user at the command line: the dataset directory (or file) and the profile HMM file.

`-dd|dataset_directory <directory name>` - Directory that contains multiple subdirectories, each one with a FASTA file of the nucleotide sequence(s) that will be used by `e-Finder`. This is the typical data structure obtained from PATRIC resource (https://www.patricbrc.org/).

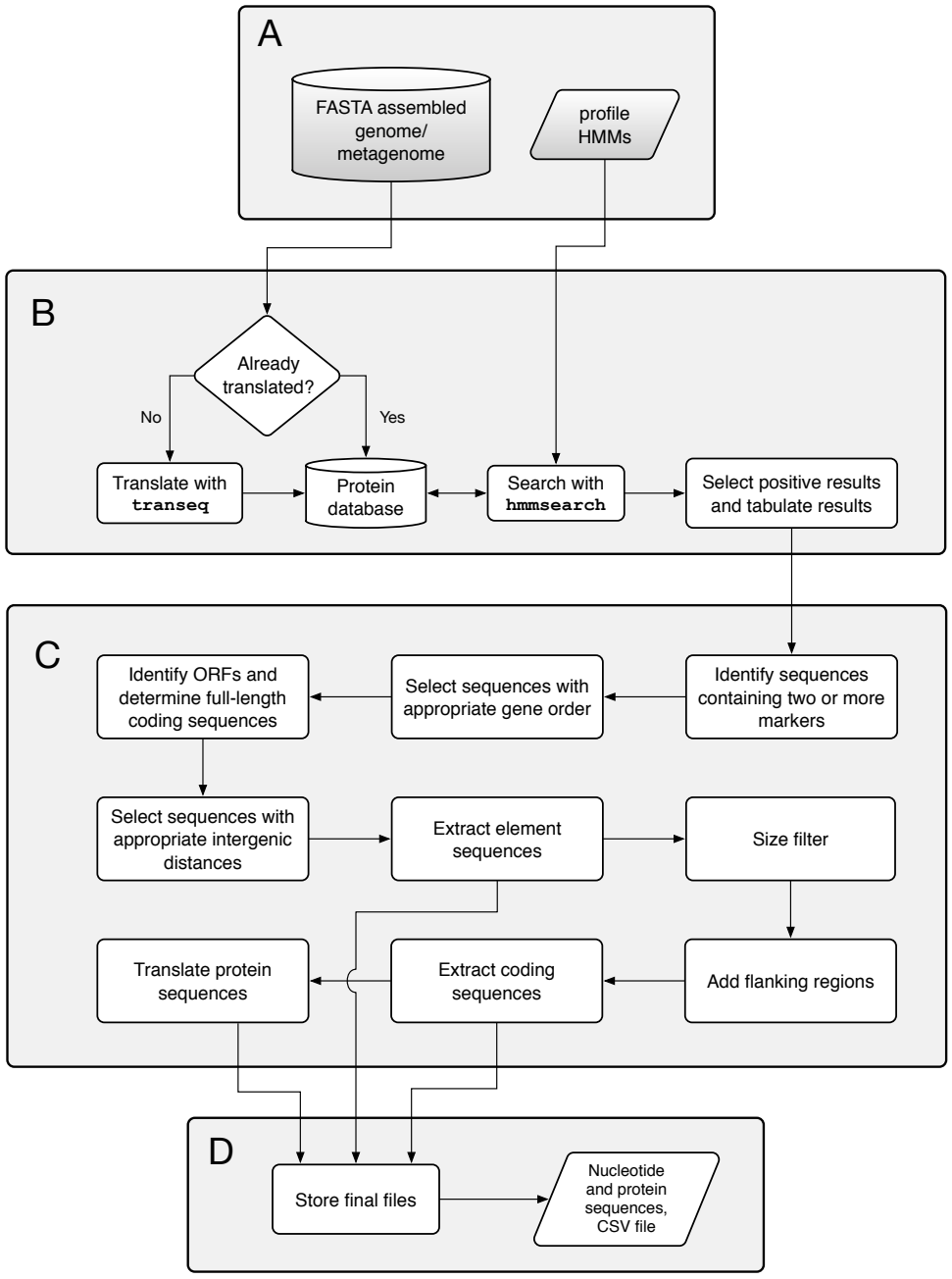

Figure 1 – Workflow of `e-Finder` program. Two types of input data are used by `e-Finder`: a set of profile HMMs and one or more FASTA files containing assembled sequences (A). In the similarity-based detection phase (B), `e-Finder` invokes `transeq` to translate the assembled sequences into the six possible open reading frames. Profile HMMs are then used as queries in similarity searches against the translated dataset using `hmmsearch`. `e-Finder` selects all sequences containing positive results, according to user-defined threshold values. In the next phase (C), `e-Finder` determines if the user-specified markers are present in the appropriate syntenic context in each of selected sequences. Regions containing the defined markers are extracted together with size-defined flanking regions and submitted to size filtering. ORFs coding for sequences detected by the models are then identified, extracted and translated into full-length protein sequences. In the final phase (D), `e-Finder` stores all element sequences and a CSV spreadsheet file listing the elements found in each assembled sequence and their corresponding markers, among other features.

`-df|dataset_file <file name>` - Dataset file in FASTA format. This file contains the nucleotide sequence(s) that will be used by `e-Finder`.

`-i|input_files <file name>` - Single or multiple files containing one or more profile HMMs each. For each marker, a distinct profile HMM file must be provided with a prefix identifier, followed by an underscore character and additional letters, or directly by an extension (e.g. `VP1_etc.hmm`, `VP2.hmm`, etc.). If using multiple files, their names must be separated by commas (e.g. `-i file1.hmm,file2.hmm,file3.hmm`). Marker prefixes can be used to define syntenic context (see parameter `-synteny`).

4.2 Optional parameters:

`-ce|circular <yes|no>` - Assume that the element is originally derived from a circular element (e.g. a prophage derived from a circular phage genome). Default: `no`.

`-conf <configuration file>` - use a configuration file that lists all parameters for execution, overriding any parameter of the command line. An example of a configuration file follows below:

```
dd=test_directory
i=cas.hmm,polB.hmm
o=output_dir
ex=fna
id=40000
fs=10000
patric_list=genome_list.csv
ol=50
gc=0
```

`-cpu|cpu_threads <integer>` - Number of threads to be used by `hmmsearch`. If not specified, `e_Finder` determines the number of threads available in the multiprocessor server and uses by default half of this value.

`-e|e-value` or `-s|score <decimal>` - E-value (`-e`) or score (`-s`) threshold value. Report `hmmsearch` hits that present values equal to or lower than the E-value or equal to or larger than the score. Only one of these parameters and the respective value shall be provided (Default = `-e 10`).

`-ex|extension <fna|faa|fasta>` - Extension for input files. When using PATRIC database data, we suggest using `fna`, since this is the extension of the contig nucleotide sequence files provided in this resource. Default: `fna`.

`-fs|flanking_size <integer>` - Size (in bp) of the 5' and 3' flanking regions that will be excised together with the multigene element (default = `5000`). If a `0` (zero) value is used, `e-Finder` will extract the sequence comprised between the start codon of the first gene and the stop codon of the last gene.

`-gc <integer>` Genetic code to define start codons and perform conceptual translation of the genes. `e-Finder` uses numbering codes defined on the NCBI Genetic Codes page and implemented on `transeq` (EMBOSS package): 0 (Standard); 1 (Standard with alternative initiation codons); 2 (Vertebrate Mitochondrial); 3 (Yeast Mitochondrial); 4 (Mold, Protozoan, Coelenterate Mitochondrial and Mycoplasma/Spiroplasma); 5 (Invertebrate Mitochondrial); 6 (Ciliate Macronuclear and Dasycladacean); 9 (Echinoderm Mitochondrial); 10 (Euplotid Nuclear); 11 (Bacterial); 12 (Alternative Yeast Nuclear); 13 (Ascidian Mitochondrial); 14 (Flatworm Mitochondrial); 15 (Blepharisma Macronuclear); 16 (Chlorophycean Mitochondrial); 21 (Trematode Mitochondrial); 22 (Scenedesmus obliquus); 23 (Thraustochytrium Mitochondrial). Default: `0`.

`-h|help` - Display help screen.

`-id|intergenic_dist <integer>`- Maximum distance (in bp) between intergenic regions (default = `5000`). This value is adopted for all intergenic distances between any pair of genes. If the `-synteny` parameter is used, its values take precedence over those set for `-id`.

`-ic|ignore_cutoff` - Ignore cutoff scores in the profile HMMs and use a custom value defined by parameters `-e` or `-s` for all input models (default = `yes`). If `-r no` is used, `e-Finder` will use the cutoff scores specified in the respective

CUTOFF SCORE tag of each profile HMM. For models not containing cutoff values, e-Finder will use the cutoff value specified by the parameter -e or -s. If none of these parameters are specified, the program will then use hmmsearch's default cutoff value (-E 10).

-mg|min_gene <integer> - Minimum number of genes (default = 2). This parameter specifies the minimum number of genes that must be found for a sequence to be considered positive. For instance, if profile HMMs from four different proteins are used and the user specifies -mg 2, any sequence containing at least two of the four markers will be considered for downstream analysis of the remaining criteria.

-o Output directory name (default: output_dir). This is the directory where e-Finder will store all output directories/file. All similarity search results are stored in the all_results subdirectory. If the user specifies an output directory that already exists, e-Finder inspects the all_results subdirectory and uses the hmmsearch results from the previous run. This feature saves processing time since it skips the relatively slow similarity search step. For each run, e-Finder creates a run_# (e.g. run_1, run_2, run_3, etc.) subdirectory where all output files are stored.

-ol|overlap <integer> - Maximum allowed overlap distance (in bp) between open reading frames in the same coding strand. If a 0 (zero) value is used (default), no overlap is allowed.

-pl|patric_list <file name> - Input file in PATRIC-like (https://www.patricbrc.org) tabular format. This two-column file lists accession codes and organism names, respectively, and provides information for e-Finder to generate a final CSV file reporting all found multigene regions associated with the respective organism names. Accepted format:

```
genome_id genome_name
1309411.5 'Deinococcus soli'
1123738.3 'Echinacea purpurea'
551115.6  'Nostoc azollae'
1856298.3 'Osedax'
```

`-sf|size_filter <integer>` - Minimum size (bp) of the excised element, not including user-defined flanking regions (parameter `-fs`). Default: `1000`.

`-synteny <string>` - Define gene order and maximum allowed distance (kb) between genes. Each marker is defined by a letter and the distances can be specified by decimals. Intergenic distances defined for parameter `-synteny` take precedence over those set for `-id`. If parameter -synteny is not specified, `e-Finder` will accept any sequences presenting the minimum number of genes (specified in parameter `-mg`), with any intergenic distances (up the maximum value defined by parameter `-id`).

If the region is assumed to be exogenously derived from a circular element (parameter `-ce yes`), then an additional distance between the last and the first marker listed must be given. In the example below, value 2.5 refers to the distance between markers `e` and `a`.

- Linear element: `a,2000,b,1500,c,3000,d,3500,e`
- Circular element: `a,2000,b,1500,c,3000,d,3500,e,2500`

`-v|version` – display program's version.

5 Running e-Finder

To be added

5.1 Understanding the parameters

To be added

6 Inspecting the output files

To be added

8 References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res. 25:3389-3402.

Eddy SR. (2011). Accelerated Profile HMM Searches. PLoS Comput Biol. 7(10):e1002195.

Krupovic M, Makarova KS, Forterre P, Prangishvili D, Koonin EV. (2014). Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biol. 12:36.

Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, Conrad N, Dietrich EM, Disz T, Gabbard JL, Gerdes S, Henry CS, Kenyon RW, Machi D, Mao C, Nordberg EK, Olsen GJ, Murphy-Olson DE, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Vonstein V, Warren A, Xia F, Yoo H, Stevens RL. (2017). Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. 45(D1): D535-D542.