e-Finder

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

A Quick Guide and Tutorial

e-Finder - A Quick Guide and Tutorial

1 Introduction

e-Finder is a generic tool for detecting and extracting multigene elements from assembled genomes using profile HMMs. e-Finder executes hmmsearch program (HMMER package) to run similarity searches using profile HMMs 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.

3.3 How to cite

- If you use this program for your publication, please cite `e-Finder` it as:

`e-Finder` program (developed by Liliane S. Oliveira and Arthur Gruber, University of São Paulo, Brazil, unpublished).

4 Understanding `e-Finder`'s workflow

4.1 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.

4.1.1 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, phages of the *Microviridae* family contain a few protein-coding genes, including those of the VP1 (major capsid protein) and VP4 (replication initiation protein). Thus, one can use profile HMMs constructed from proteins coded by these genes to fish prophages in bacterial genomes. If multiple profile HMMs are available for each protein, they can 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 as filters in the similarity search step (see below for more details).

4.1.2 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 contains 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.

4.2 Detection of candidate sequences

`e-Finder` uses profile HMMs as a 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 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 an alternative 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 the `e-Finder` detects a previous similarity search file, it will only analyze that file to retrieve the results that fall within 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. All models from the `Viral MinionDB`

(http://www.bioinfovir.icb.usp.br/miniondb) are provided with the appropriate cutoff scores. `E-Finder` executes `hmmsearch` with the `-T` option using the respective cutoff value.

4.3 Selection of positive sequences

Once the positive sequences have been detected, `e-Finder` checks whether the user-defined criteria for the synthetic context have been met (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.

4.4 Sequence extraction and report generation

In the final phase (Figure 1D), `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.

5 Understanding `e-Finder` parameters

5.1 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 the `PATRIC` repository (https://www.patricbrc.org/).

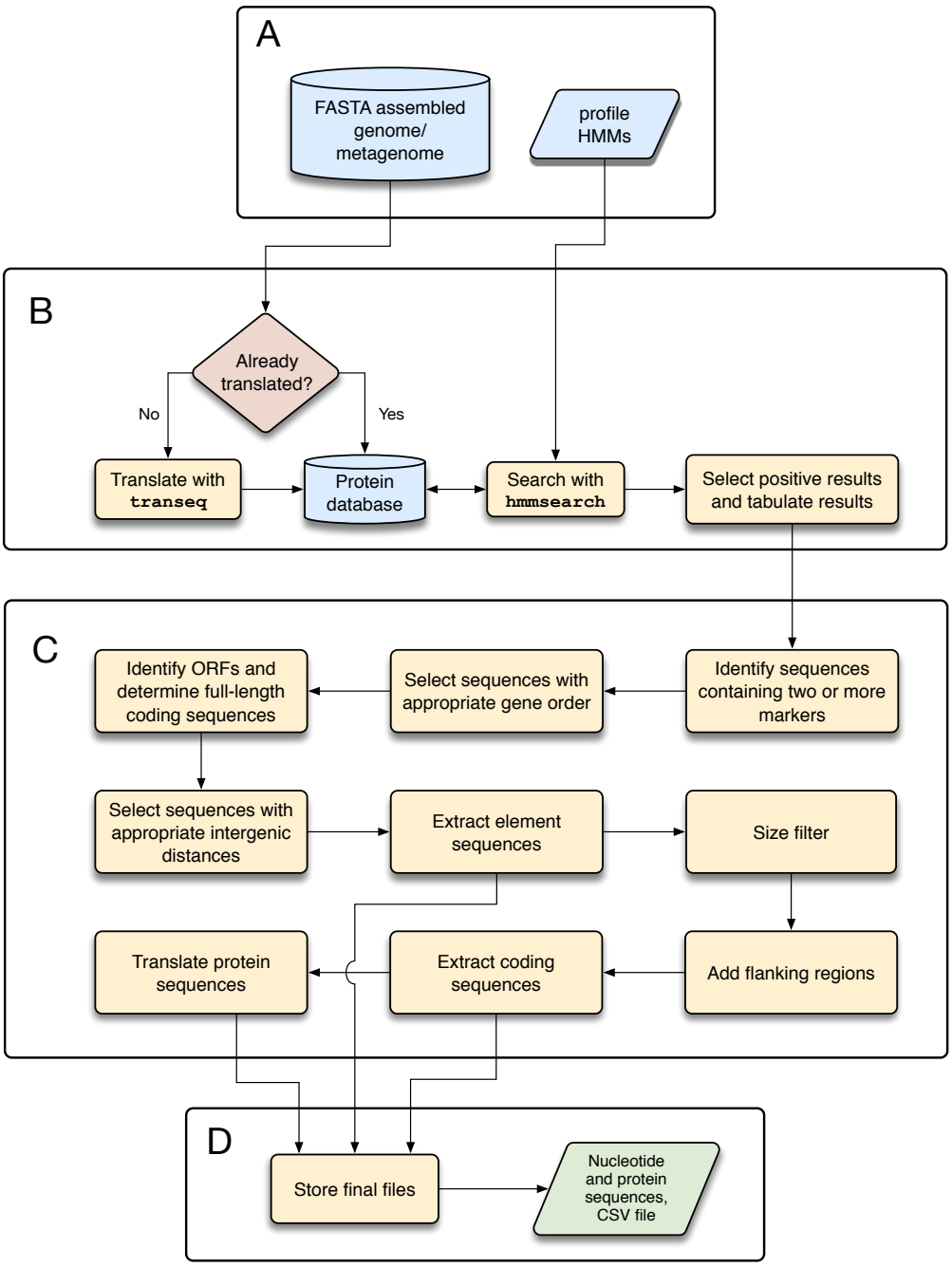

Figure 1 – Workflow of `e-Finder`. Two types of input data are used by `e-Finder`: a set of profile HMMs and one or more FASTA files containing assembled sequences (A). In the similarity-based detection phase (B), `e-Finder` invokes `transeq` to translate the assembled sequences into the six possible 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 whether 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 is 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. `prot1.hmm`, `prot2.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`).

5.2 Optional parameters:

`-ce|circular <yes|no>` - Assume that the element is originally derived from a circular element (e.g. 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=gene1.hmm,gene2.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`).

-ed|element_distance <integer> - Minimum distance between elements. Default = 5000.

-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 extracts 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.

-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).

-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`.

`-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. The original multi-column file distributed by PATRIC can be used or, alternatively a simplified version. In this case, a 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. Simplified 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.

6 Running `e-Finder` to find prophages of the *Alpavirinae* subfamily (*Microviridae* family)

In this tutorial we will use two profile HMMs built from multiple sequence alignments of VP1 and VP4 protein sequences derived from phages of the subfamily *Alpavirinae* (*Microviridae* family). Protein sequences of VP1 and VP4 were obtained from Roux *et al*. (2012). We will use these models to interrogate six assembled bacterial genomes obtained from the `PATRIC` database in order to find prophages of these prokaryotic viruses.

To install the data for the tutorials, copy the file `tutorial.tar.gz` to a directory of your choice. Decompress the file using the following command:

```
tar xzvf tutorial.tar.gz
```

This command will create the `tutorial` directory. This directory contains the following items:

- Subdirectory `hmms` - contains two profile HMM files:
  - `VP1-Alpa.hmm` and `VP4-Alpa.hmm` – these files correspond to profile HMMs constructed from VP1 and VP4 protein sequences, respectively. The protein sequences were derived from *Alpavirinae* phages (*Microviridae* family) described by Roux *et al.* (2012).
- Subdirectory `patric` – contains:
  - Six subdirectories corresponding to bacterial genome data obtained from the `PATRIC` database (Wattam *et al*., 2017) with its typical data structure: `1262921.3`, `1313.13787`, `1898203.1819`, `46170.608`, `483216.6` and `93974.3`.
  - `patric_list.csv` – this is a tabular file listing accession codes and organism names, among other data, that provides information for `e-Finder` to generate a final CSV file reporting all found multigene regions associated with the respective organism names.
- Subdirectory `results` - contains multiple subdirectories with pre-run results of this tutorial.
- `alpavirinae.cnf` – a configuration file to run this tutorial.

To execute `e-Finder`, the configuration file can be used:

```
e-finder.pl -conf alpavirinae.cnf
```

The configuration file content follows below:

```
dd=patric
i=hmms/VP1-Alpa.hmm,hmms/VP4-Alpa.hmm
o=alpavirinae_dir
ic=no
ex=fna
id=3000
fs=5000
patric_list=patric/patric_list.csv
ol=50
gc=11
sf=1000
```

- Alternatively, instead of running `e-Finder` with the configuration file, the program can be executed with all parameters specified in a line command:

```
e-finder.pl -dd patric -i hmms/VP1-Alpa.hmm,hmms/VP4-Alpa.hmm
-o alpavirinae_dir -ic no -ex fna -id 3000 -fs 5000 -pl
patric/patric_list.csv -ol 50 -gc 11 -sf 1000
```

7 Inspecting `e-Finder`'s output files

Once `e-Finder` finishes the processing, an output directory is created with several files and subdirectories:

- `error.log`: this file reports any possible error that occurred during execution. If no error occurred, this file will be empty.
- `alpavirinae_dir`: the output subdirectory, as specified in the configuration file. A `results` subdirectory is also provided, with pre-run results. The following subdirectories are present:
    - `all_results`: all similarity search results are stored in this directory. Since all subsequent analyses use the `hmmsearch` results as a starting point, it is possible to re-run `e-Finder` using different parameters (minimum number of genes, intergenic distances, etc.). In this case, the user just needs to execute `e-Finder` in a directory where `all_results` folder already exists. If the user wants to use other profile HMMs, then a new directory must be created to execute a clean run. Subdirectories containing the similarity search results of each analyzed genome are stored with names corresponding to the respective genome IDs. The results of each genome are stored separately for each protein according to the names of the profile HMM files (in this case `VP1-Alpa` and `VP4-Alpa`).
    - `run_1`: this folder contains the following files and subdirectories:
        - `logfile.txt` – a log file reporting all steps of `e-Finder`'s execution.
        - `final_report.csv`: this is a tab-delimited result file that summarizes the results for all elements found. Stored information includes organism name, contig size, element coordinates in the contigs, gene coordinates and orientation, intergenic distances, etc. This file can be imported into MS Excel to facilitate inspection.

- `selected`: this is the directory where `e-Finder` stores the results of all selected (positive) datasets, each one in a specific subdirectory. In the case of this tutorial, three genomes are positive and present within this directory: `1262921.3`, `483216.6` and `93974.3`. If a `PATRIC`-like data structure is used, the subdirectories are named with the corresponding identifier. Each subdirectory contains some files, including nucleotide sequences of the positive contig and ORFs of the proteins that were positive to the profile HMMs used in the search. Sequences of the corresponding proteins are also stored. The file `elements.fasta` contains the nucleotide sequences of all elements found in the positive datasets. Additional files include protein sequences conceptually translated from the elements, in this tutorial they are named `VP1-Alpa_proteins.fasta` and `VP4-Alpa_proteins.fasta`. These are the proteins detected by the profile HMMs.
- `discarded`: this is a directory containing the results of all positive datasets that did not fulfill all criteria to be selected. For instance, sequences that were positive for only marker, or whose intergenic length was out of the maximum defined distance.

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.

Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F & Gordon JI. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466(7304):334-338.

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.