HMM-Prospector

A tool for surveying genomic and metagenomic data with profile HMMs

A Quick Guide

HMM-Prospector - A Quick Guide and Tutorial

1 Introduction

HMM-Prospector is a Perl script that uses a single or multiple profile HMM as a query in similarity searches against a FASTQ/FASTA dataset using hmmsearch program (Figure 1). HMM-Prospector processes the results and generates tabular files with qualitative and quantitative results. Previously run hmmsearch result files (short tabular format) can also be used as datasets. If models from the vFam database (Skewes-Cox *et al.*, 2014) are used, then HMM-Prospector can add taxonomic information to the result files. HMM-Prospector helps you to conveniently answer the following questions:

- Which profile HMMs are most recognized by the dataset reads?
- How many reads of the dataset are detected by a profile HMM?
- Based on the detection by profile HMMs, which viral families (for vFam only) are most represented in the dataset?

2 Main features

- Fully configurable
- FASTQ and FASTA files can be used as input
- Previously run hmmsearch results can be reprocessed with new cutoff values
- Score and e-value cutoff values can be used
- Taxonomic data is incorporated in the results when using vFam models

3 Before using **HMM-Prospector**

3.1 System requirements

HMM-Prospector 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).

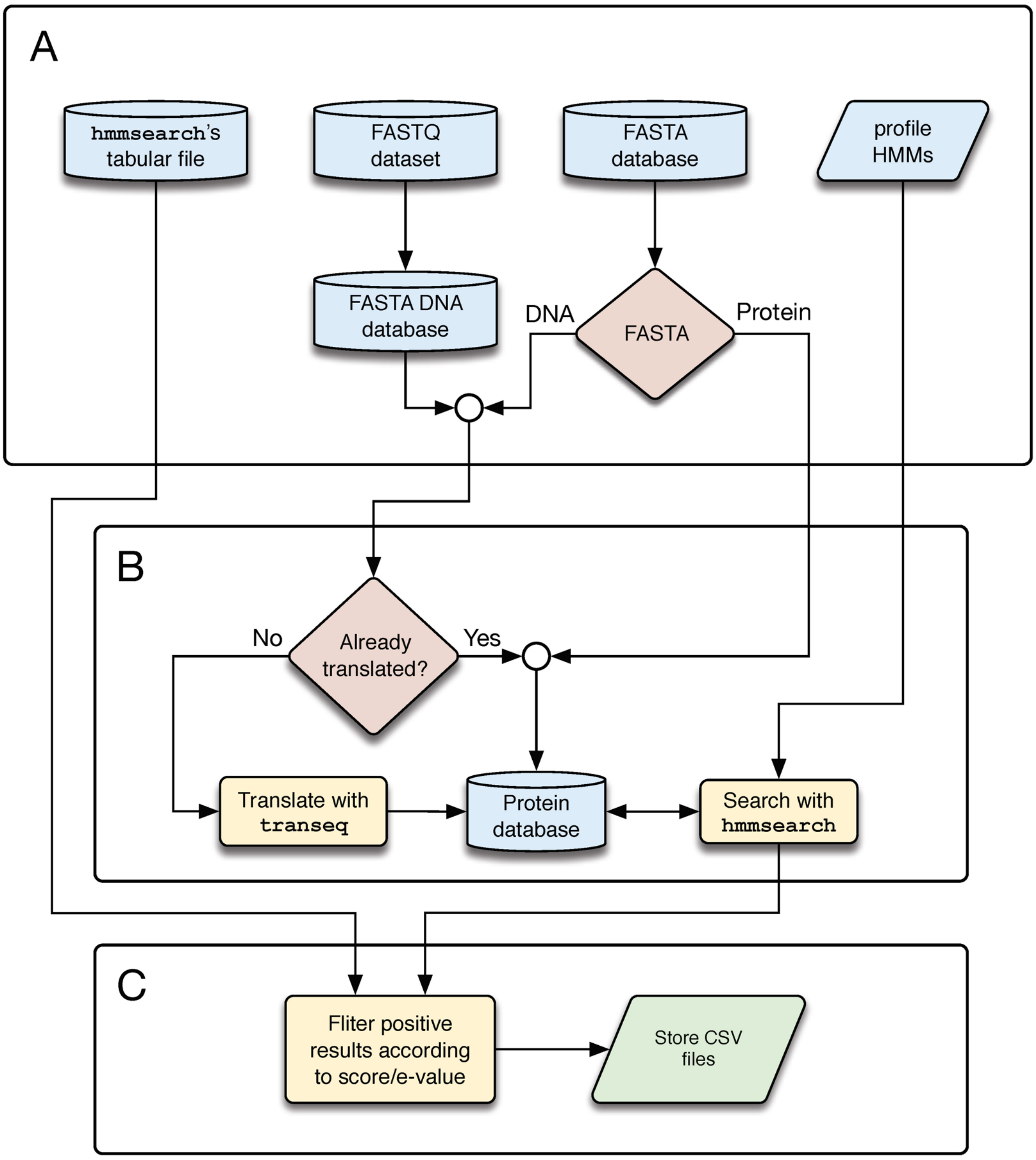

Figure 1 – Workflow of `HMM-Prospector` program. The input consists of a profile HMM file and a dataset in either FASTQ or FASTA (DNA or protein sequences) formats. A pre-run `hmmsearch`'s tabular result file is also accepted (A). If necessary, `HMM-Prospector` invokes `transeq` to translate the sequences into the six possible open reading frames. Profile HMMs are then used as queries in similarity searches against the translated dataset using `hmmsearch` (B). In the next phase (C), `HMM-Prospector` lists all sequences containing positive results, according to user-defined cutoff values (score or e-value) and stores all results into CSV spreadsheet files.

3.2 Third-party programs

HMM-Prospector requires the following programs and databases:

- hmmsearch (HMMER3 package - http://hmmer.org/). This program is used to run similarity searches using profile HMMs against metagenomic datasets. The program must be located in a directory listed in the PATH of the operating system.
- transeq (EMBOSS package - http://emboss.sourceforge.net/). This program is used to translate nucleotide sequences into the six possible frames.
- fastq_to_fasta – (FASTX-Toolkit – http://hannonlab.cshl.edu/fastx_toolkit/)
- vFam database (Skewes-Cox *et al*, 2014) – This database is only required if you want to analyze metagenomic data using vFam models. vFam can be downloaded from http://derisilab.ucsf.edu/software/vFam/. You will need the respective model file (vFam-A or vFam-B) and annotations files.

3.3 How to cite

- If you use this program for your publication, please cite HMM-Prospector it as:

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

4 Understanding **HMM-Prospector** parameters

4.1 Mandatory parameters:

HMM-Prospector has one mandatory parameter that must be specified by the user at the command line:

-d <file name> Dataset file (FASTQ, FASTA or hmmsearch's tabular output file are accepted). For FASTA files, both DNA and protein sequence datasets can be used.

4.2 Optional parameters:

-a `<directory name>` Directory containing profile HMM annotations (valid only when using `vFam` models as input)

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

-e `<decimal>` or `-s <decimal>` E-value (`-e`) or score (`-s`) cutoff value. Report `hmmsearch` hits that present values equal to or lower than E-value or equal to or larger than score. One parameter and the respective value must be provided. If an `hmmsearch` tabular result file is used as input, then parameters `-e` or `-s` become mandatory.

-h or `help` Show help screen.

-i `<file name>` Input file (single or multiple profile HMMs) – this parameter is ignored when an `hmmsearch`'s tab-delimited result file is used as input, but is **mandatory** when using a FASTA or FASTQ dataset.

-o Output directory name (default: `output_dir`).

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

-v Display program version.

5 Running **HMM-Prospector**

In this tutorial we will execute `HMM-Prospector` using a set of 15 profile HMMs with specificities to phages of the family *Microviridae*. We will also use a viral metagenomic dataset from human fecal samples, described by Reyes *et al.* (2010). Type the command below:

```
hmm-prospector.pl -d dataset.fastq -i profiles.hmm -cpu 20 -r
no -o result_dir &
```

5.1 Understanding the parameters

The command above executes `HMM-Prospector` using the file `dataset.fastq` as the dataset and `profiles.hmm` as an input file containing multiple profile HMMs. `HMM-Prospector` invokes `hmmsearch` with 20 threads (parameter `-cpu 20`). All result files are stored in the `result_dir` directory.

`HMM-Prospector` automatically detects a `CUTOFF SCORE` tag in the profile HMM. Cutoff scores are not available on `vFam` models or any other conventional profile HMM. The `CUTOFF SCORE` tag can be manually inserted by the user (e.g. `CUTOFF SCORE 35`). All models from the `Viral MinionDB` (http://www.bioinfovir.icb.usp.br/miniondb) are provided with the appropriate cutoff scores. `HMM-Prospector` executes `hmmsearch` with the `-T` option using the respective cutoff value. The purpose of using cutoff scores is to increase the specificity of the survey, especially in large metagenomic data. Nevertheless, using cutoff scores is optional. The command above included the parameter `-r no`, which informs the program to use cutoff scores specified in the respective `CUTOFF SCORE` tag of each profile HMM. Instead of using the parameter `-r no`, `HMM-Prospector` can be executed with the parameter `-s`. For instance, `-s 20` specifies a score cutoff of `20` to display positive hits.

5.2 Inspecting the output files

Once `HMM-Prospector` finishes the processing, an output directory is created. Within this directory, you will find several files and subdirectories :

- `file.log`: a log file created by `HMM-Prospector` that reports all steps of the execution. The command used for `hmmsearch` is also displayed.
- `error.log`: a log file created by `HMM-Prospector` that stores any error message. If the execution proceeds without errors, this file remains empty.

- `prefix_hmmsearch.txt`: this is a full result file generated by `hmmsearch` with full alignments (output data when using `hmmsearch` with `-o` parameter). The prefix corresponds to the name of the dataset file.
- `prefix_hmmsearch.tab`: this is a tab-delimited result file generated by `hmmsearch` (output data when using `hmmsearch` with `--tblout` parameter). The prefix corresponds to the name of the dataset file.
- `table1.csv`: this CSV file lists the positive sequences according to each tested profile HMM. The table below shows an example of the output.

Target Name	query_pHMM	E-value	Score	Description of target
SRR073436.48099_2	microviridae_3_861-880	4.8e-07	37.1	FPK2RZS01E2O8V length=284
SRR073436.56819_6	microviridae_3_861-880	1.1e-06	36.0	FPK2RZS02NAOFU length=277
SRR073436.31249_5	microviridae_3_861-880	1.1e-06	36.0	FPK2RZS02MV4X1 length=260
SRR073436.70846_5	microviridae_3_861-880	1.2e-06	35.8	FPK2RZS02M6J8L length=276
SRR073436.44004_4	microviridae_3_861-880	1.3e-06	35.7	FPK2RZS01CS6F6 length=277
SRR073432.17523_2	microviridae_3_861-880	1.4e-06	35.7	E3MCXE401AUQXA length=272
SRR073436.66730_3	microviridae_3_861-880	1.5e-06	35.5	FPK2RZS02MIOFQ length=279

- `table2.csv`: this CSV file presents a list of all profile HMMs that detected positive reads in the sequencing dataset. The respective numbers of positive reads are listed in the right column.

HMM	# of seqs
Alpavirinae_10_1211-1226	355
Alpavirinae_2_85-105	646
Alpavirinae_8_993-1012	1038
Gokushovirinae_10_1246-1263	293
Gokushovirinae_1_41-60	262
Gokushovirinae_2_86-105	323
Gokushovirinae_3_66-84	275
Gokushovirinae_7_974-994	321
Gokushovirinae_8_1080-1100	305
Microviridae_1_101-119	627

Microviridae_2_486-505	1368

6 Running **HMM-Prospector** with **vFAM** models

HMM-Prospector can be used to perform similarity searches against sequencing datasets using vFam models and their corresponding annotation files. The command below shows how to execute HMM-Prospector using vFam models, where annotation_files is the directory that contains the annotation files provided by the vFam repository:

```
hmm-prospector.pl -d dataset.fastq -s 10 -a /your_directory_paths/annotationFiles -cpu 20 -i vfams.hmm -o result_dir
```

The output files are then customized to vFam. In this case, vFam_table1.csv is produced instead of the table1.csv file and lists the positive sequences, vFam IDs, e-value and score of the respective alignments, number of sequences originally used to build the vFam model and the viral family of these sequences.

Target Name	query _pHMM	E-value	Score	# of seqs	Family
AAUC0:1:1114:14372:20845_1	vFam_ 3011	9.00E-53	186.4	3	Retroviridae
AAUC0:1:2114:3987:18752_6	vFam_ 3011	1.80E-51	182.1	3	Retroviridae
AAUC0:1:1112:10162:8600_4	vFam_ 3011	6.30E-51	180.3	3	Retroviridae
AAUC0:1:1113:14155:12256_2	vFam_ 717	1.30E-23	89.5	12	Picornaviridae
AAUC0:1:1111:18729:20232_2	vFam_ 497	2.10E-21	81.9	10	Bunyaviridae
AAUC0:1:1111:17714:9335_4	vFam_ 3011	5.20E-21	81.5	3	Retroviridae
AAUC0:1:1104:16536:7046_4	vFam_ 3011	5.50E-21	81.4	3	Retroviridae
AAUC0:1:1112:3051:11448_2	vFam_ 3011	8.50E-21	80.7	3	Retroviridae
AAUC0:1:1111:18729:20232_5	vFam_ 497	6.50E-21	80.3	10	Bunyaviridae

7 Rerunning **HMM-Prospector** on previously run **hmmsearch's** results

HMM-Prospector selects hmmsearch's results, according to a user-defined cutoff score or e-value and stores the information in a CSV file. In case the user wishes to employ another cutoff, it is not necessary to run HMM-Prospector in the similarity search mode (when it invokes hmmsearch). If a run has been previously performed, and a tab-delimited result file is available, then this file can be used as an input. The user may run HMM-Prospector in selection mode (when it only selects results from previous runs) using this file and specifying a different score or e-value cutoff. Instead of invoking hmmsearch again, HMM-Prospector saves time and just parses out the hmmsearch's tab-delimited result file and stores the data into new CSV files. In the example below, the result file generated in the previous command (see item 5) is used as an input for a new execution, with an e-value cutoff of 10^{-5}:

```
hmm-prospector.pl -d prefix_hmmsearch.tab -e 0.00001 -o new_results_dir
```

8 Using profile HMM seeds with GenSeed-HMM

Once you identify the most promising datasets to find viral sequences of your interest, you can use a set of selected models as seeds for targeted progressive assembly with GenSeed-HMM program (Alves *et al.*, 2016). GenSeed-HMM is publicly available at https://sourceforge.net/projects/genseedhmm/. The program is fully documented and includes a tutorial.

9 References

Alves JM, de Oliveira AL, Sandberg TO, Moreno-Gallego JL, de Toledo MA, de Moura EM, Oliveira LS, Durham AM, Mehnert DU, Zanotto PM, Reyes A & Gruber A. (2016) GenSeed-HMM: A Tool for Progressive Assembly Using Profile HMMs as Seeds and its Application in *Alpavirinae* Viral Discovery from Metagenomic Data. Front. Microbiol. 7: 269.

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.

Skewes-Cox P, Sharpton TJ, Pollard KS & DeRisi JL. Profile hidden Markov models for the detection of viruses within metagenomic sequence data. PLoS One. 9(8):e105067.