# User Guide for `HmmCleaner`

Arnaud Di Franco [arnaud.difranco@sete.cnrs.fr]

Version 0.1 / Feb 13, 2018

## Contents



## 1 Aim and features

The aim of `HmmCleaner` is to clean multiple sequence alignments (MSA) by removing Low Similarity Segments (LSS) that could correspond to sequence errors. The objective of removing LSS is to avoid production of erroneous signal while performing subsequent analysis on MSA.

`HmmCleaner` handle MSA in Fasta and MUST Ali format and can output files in both format. It also outputs the list of blocks removed for each sequence of the MSA as well as the score alignment of each sequence to the profile HMM.

`HmmCleaner` works using a scoring matrix of 4 cost parameters (c1 < c2 < c3 < c4) that can be modified by users either through the selection of a predefined set with the `large` and `specificity` options or by manually choosing their own values with the `costs` option.

`HmmCleaner` is dependant of `HMMER` version 3.1b2 available at http://hmmer.org. All executable from `HMMER` have to be present in the $PATH variable of users.

## 2 Functional overview

A graphical overview of `HmmCleaner`'s pipeline is available at next page (Figure 1).

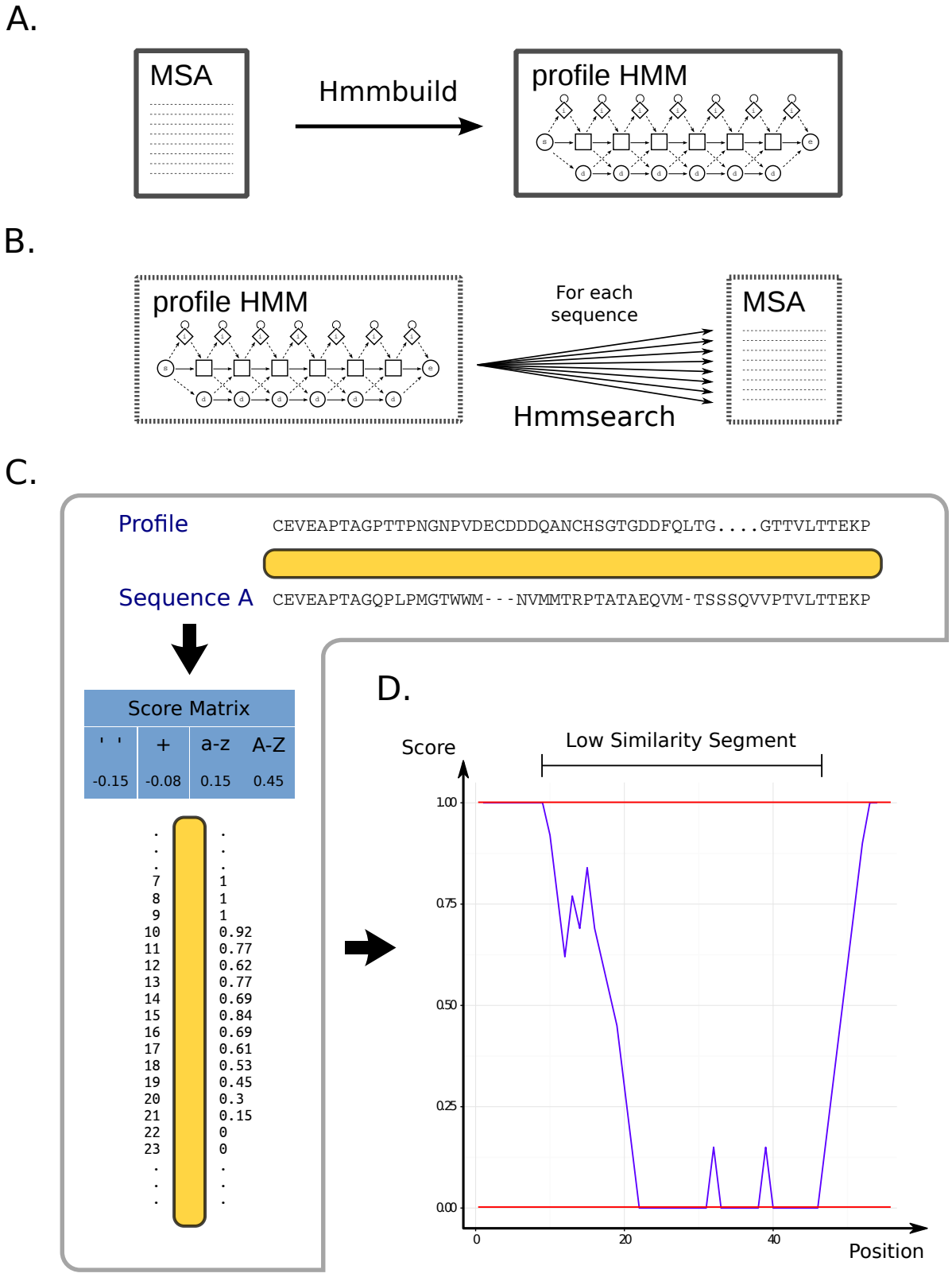


Figure 1: Overview of `HmmCleaner`'s pipeline (see text for details).

### 2.1 Profile creation

`HmmCleaner` detects low similarity segments (LSS) through four steps. First, a pHMM is built from the MSA using `HMMER` (Figure 1A). This pHMM can be built upon either (i) all sequences of the MSA (complete strategy) or (ii) all sequences excepted the currently analyzed one (leave-one-out strategy). Users can affect this step with the `profile` option.

### 2.2 Similarity search

Second, each sequence of the MSA is evaluated with the pHMM (Figure 1B), which yields profile-sequence alignments.

### 2.3 Score analysis

Third, the analysis of each profile-sequence alignment is based on the four discrete categories of column-wise probabilities provided by `HMMER`. The two first categories represent residues that fit poorly to the pHMM: blank character (null probability, parameter c1) and '+' character (low probability, parameter c2). In opposition, the two last last categories represent residues that fit to the pHMM: amino acid characters in lower case (good probability, parameter c3) and upper case (high probability, parameter c4). A cumulative similarity score increases when the residue is expected from the profile or decreases it otherwise (Figure 1C). Parameters c1 and c2 are therefore negative and parameters c3 and c4 positive. The cumulative score is computed from left to right starting with a value of 1. Its value is strictly restricted between 0 and 1 included. An LSS start at the last position with a cumulative score of 1 when this one reaches a null value. Its end is defined by the last position with a null value once the cumulative score goes back to 1 or when the end of the sequence is reached (Figure 1D).

## 3 Outputs

`HmmCleaner` outputs 3 types of files named regarding the input MSA file with the `_hmm` suffix. The first file is the cleaned MSA file in Fasta format (default) or in MUST ali format (`ali` option). The second file is the log file. It includes the list of low similariry segments (LSS) removed for each sequence of the MSA. Finally, the score file gives the alignment between the original sequence, the score observed by `HMMER` and the output sequence. Users can decide to retrieve only the LSS detected with the `log-only` option.

## 4 Annexes

### 4.1 Command line interface

```
USAGE
    HmmCleaner.pl <infiles> [options]

REQUIRED ARGUMENTS
```

```
    <infiles>
        list of alignment file to check with HmmCleaner

OPTIONS
    -costs <c1> <c2> <c3> <c4>
        Cost parameters that defines the low similarity segments detected by
        HmmCleaner. Default values are -0.15, -0.08, 0.15, 0.45 Users can
        change each value but they have to be in increasing order. c1 < c2 <
        0 < c3 < c4 Predefine value are also available with --large and
        --specificity options but user defined costs will be prioritary if
        present.

    --changeID
        Determine if output will have defline with generic suffix
        (_hmmcleaned)

    --noX
        Convert X characters to gaps that will not be taken into account by
        HmmCleaner.

    -profile=<profile>
        Determine how the profile will be create complete or leave-one-out
        (default: complete) leave-one-out = without the analyzed sequence
        (new profile each time) complete = all sequences (same profile for
        each sequence) First case is more sensitive but need more ressources
        (hence more time)

    --large
        Load predifined cost parameters optimized for MSA with at least 50
        sequences. Can be use with --specificity option. User defined costs
        will be prioritary if present.

    --specificity
        Load predifined cost parameters optimized to give more weigth on
        specificity. Can be use with --large option. User defined costs will
        be prioritary if present.

    --log_only
        Only outputs list of segments removed.

    --ali
        Outputs result file(s) in ali MUST format.

    -v[erbosity]=<level>
        Verbosity level for logging to STDERR [default: 0]. Available levels
        range from 0 to 5.

    --version
```

```
--usage
--help
--man
    Print the usual program information
```