Manual HMMERCTTER Version 1

Table of Contents

Introduction

HMMERCTTER is a decision support system for the classification of protein superfamily sequences with a reliable cut-off, based on iterative clustering and classification procedures with clusters that show 100% precision and recall. Strictly it is neither clustering not classification that is performed but it can quite safely be envisaged as a two-step procedure of training (clustering) and target analysis (classification). Hence, HMMERCTTER consists of two separate pipelines, currently written in Matlab. Besides that it depends on HMMER, it requires MAFTT for multiple sequence alignment, Dendroscope for tree handling and it has a number of PERL scripts. Not only does this imply HMMER-CTTER has a lot of dependencies, it also correctly suggests it performs a complex task and its usage can be complicated. Please use this manual to guide your HMMERCTTER studies. It is very powerful albeit that we cannot foresee the complexity of all superfamilies. We therefore appreciate all your feedback. Positive feedback will smooth our egos but we do prefer negative feedback since this we can use to remove possible bugs and to improve algorithms and pipeline.

Dependencies

HMMERCTTER was developed on Linux, it will likely run on a Mac and possibly on Windows using Cygwin. This has NOT been tested. Most dependencies can be installed from your Linux repos but do check you have the correct version, particularly for HMMER since HMMER v2 is not compatible with v3. In general higher versions will not cause any problems.

1 Matlab v7.12.0: Matlab is not free software but a license is not expensive and you can get a 30 day trial at: https://www.mathworks.com/campaigns/products/ppc/google/matlab-trial-request.html?s_eid=ppc_29850150442&q=matlab
2 MAFFT v7: http://mafft.cbrc.jp/alignment/software/
3 HMMER v3: http://hmmer.org/download.html
4 Dendroscope v3:
http://ab.inf.uni-tuebingen.de/data/software/dendroscope3/download/welcome.html
5 PERL v5.6.1: https://www.perl.org/
6 BioPerl v1.6: http://bioperl.org/INSTALL.html

Installation

HMMERCTTER comes as a tarballed gzip that can be downloaded from the XXXX repository facility. It contains a collection of Matlab and PERL scripts and a folder with libraries that are required. Once you have installed the dependencies all you should do is extract the archives and you are good to go. Some of you might experience problems since communication between the dependencies sometimes need certain libraries that may or may not have been installed on your machine. If you do encounter a problem, first check if the dependencies are functioning independently of HMMERCTTER. If so check out the FAQ. Please report any other problem, we will see if we can solve it and add it to the FAQ.

Preparation of your datafiles

As you might be aware, the format problem is a difficult problem and this section describes how to format your data in order to prevent any problems. You will need three datasets. For the training you need a high quality phylogeny and 100% corresponding sequence dataset. For the target analysis you will need a set of sequences to be tested. Sequence files should be in fasta format that consists of a sequence annotation line that starts with a ">" and the sequence annotation on the same line, followed by the sequence on the next line(s). Both sequence and annotation can yield problems. MAFFT is run using the option "--anysymbol" that should prevent problems but be aware of the fact that sequences with uncommon symbols as U and Z can show strange behavior. Then we kindly suggest to pass the sequence set through a PERL script (Folder Additional) that will change your sequence names to a number, meanwhile generating an index file, such that at any time you can change the code to the original code! We strongly advise this in order to prevent problems with strange symbols that sometimes generate problem or sometimes are changed by a dependency. This would be fatal since the sequence annotation is crucial for the process.

Once you have an impeccable sequence file, proceed to the alignment and phylogeny. HMMERCTTER uses MAFFT for its alignments and this is fine since it concerns subfamilies. Superfamily alignment is a different business and is work for experts, both in the sense of the biological system knowledge as the alignment process. Prior to phylogeny, the MSA should be trimmed in order to

remove unreliably aligned columns. As a very rough guideline, the reference sequence in your trimmed MSA should contain in between 30 and 70% of its residues (that is when it concerns a single domain protein). Maximum Likelihood is the recommended tree m,ethod. Do check the reliability of your tree with statistical support and by introducing several outliers. Sequences with low support that jump clades will likely results in poor classification. Comparison of trees with different outgroups con identify sequences that generate conflicts. These are typically sequences that end up at large distances and you might want to consider removing these sequences. This does indeed result in a bias but on the bright side: these sequences are often pseudogenes. Trees should be in Newick format.

Regarding the target set: we strongly advise to provide a dataset that contains homologs only. Using large datasets results in long running time but moreover, a set with homologs only will yield results that make it easier to understand how high your coverage is.

Note that although you need an alignment for the phylogenetic reconstruction, you do not use this alignment in HMMERCTTER since HMMERCTTER aligns the subfamilies. Hence, you provide:

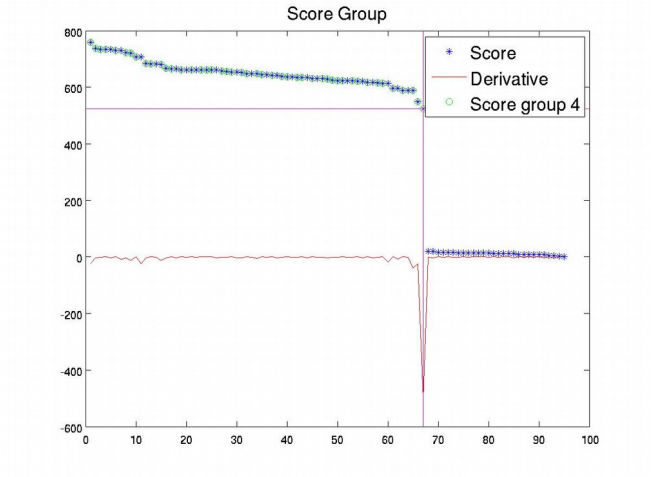
1 Train.fsa: Your set of UNALIGNED training sequences
2 Train.tree: Your tree in Newick format with or without statistical support
3 Target.fsa: the sequences you want to classify.

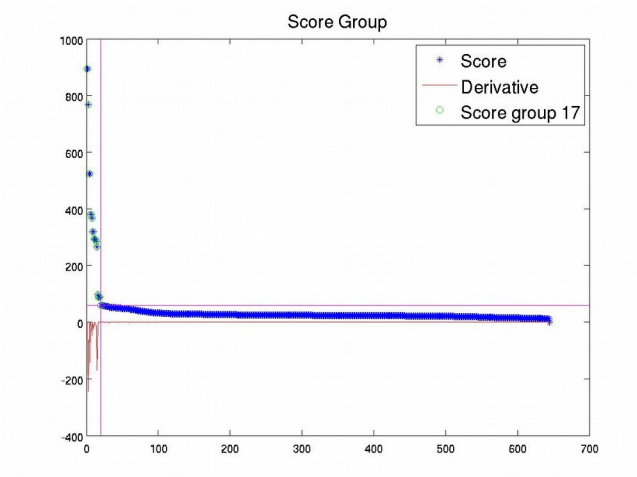
Running HMMERCTTER

Phase 1 Training

Open Matlab and change to the directory where you extracted the HMMERCTTER package. Type "Clustering" on the prompt and off you go. HMMERCTTER asks for both the tree and sequence file that form the training set, it also asks you for a folder where you want to put the output. Always define a novel empty folder. In addition you can set the minimal groupsize, which is set at 8 by default. We recommend not to accept groups smaller than 3 sequences.

HMMERCTTER will search for the largest monophyletic cluster with 100% P&R and will ask you if you accept this group. Here it becomes clear why we refer to HMMERCTTER as expert decision support system. Often clusters will be larger than functional clusters the expert is familiar with. Hence, the expert can refuse clusters that show 100% P&R. Then, HMMERCTTER provides visual information: the tree with the group in question indicated in red and the corresponding *hmmsearch* plot. Below you see graphs of a good group (4) and a bad group (17). Note that the score-drop is indicated by the derivative, printed in red. Be aware that good or bad is the result of a complex calculation that basically scores similarity rather than a measure based on tree distances. Often these correspond and we can state in general that high compactness and separateness of groups yield good groups (that is with good classification behavior). The next version of HMMERCTTER will contain one or more actual measures in order to provide better support in supervised clustering.





Once you accept a group, HMMERCTTER searches for the next largest 100% P&R group. Once no more groups are identified, groups are saved, a final dendrogram with colors per clade is made. The colors used are shown in Table 1.

Table 1: Colorscheme followed by HMMERCTTER clustering

Group	Description	Red	Green	Blue
G1	Red	255	0	0
G2	Blue	0	0	255
G3	Lime	0	255	0
G4	Magenta	255	0	255
G5	Cyan	0	255	255
G6	Silver	192	192	192
G7	Gray	128	128	128
G8	Maroon	128	0	0
G9	Olive	128	128	0
G10	Green	0	128	0
G11	Purple	128	0	128
G12	Teal	0	128	128
G13	Orange	255	69	0
G14	Gold	255	215	0
G15	Salmon	255	128	114
G16	Indigo	75	0	130
G17	Saddle Brown	139	69	18
G18	Deep Pink	255	20	147
G19	Midnight Blue	25	25	112
G20	Blu Violet	138	43	226

Subsequently, random colors will be selected. In addition a file is generated that contains the sequences that are not accepted, the orphans.

As is made clear in the PG case of the first manuscript, it is advisable to test various different clusterings, high classification coverage is a good measure for clustering.

Phase 2: Classification

Type "Classification" on the prompt and off you go. HMMERCTTER will ask for the folder where you stored the clustering output, the target sequence set and a folder where you want to send the classification output. Always define novel empty folders. Then it starts the automated classification phase that, depending on the size of the dataset (and your computational power) might take up to hours. Please be patient, HMMERCTTER will come back once the automated phase has finished.

The interactive phase is generally fast since only few iterations are required. In the interactive phase, sequences with scores below the threshold are tested. There are several options, "+1" and "manual selection" being the most powerful. For the manual selection you need to lower the threshold with a mouse and click Manual Selection. We advise to always add the maximal amount of sequences (using iterations or not). This maximal amount is determined by three factors. First sequences already classified will not be accepted and you must accept the group as it stood. Than, sometimes a sequence is added that results in a group that is no longer 100% P&R, which is also NOT accepted. Those are objective measures. Then, a third and unfortunately subjective measure is the threshold. High specificity is obtained when one remains a significant score-drop below the threshold. This is however not always possible (the automated phase only takes into account the 100% P&R rule) or even advisable. We advise to run the classification twice and to use the first run for orientation. The "previous" button can be used to return to the former state and there is also an "original" button.

FAQ

1 How do I know my classification is correct?
-You don't. HMMERCTTER makes predictions albeit its performance is excellent. If you really want to answer this question you need to run a phylogeny and be aware of the question: Which is correct? The phylogeny or the HMMERCTTER clustering? There is no easy answer to this question.

2 Can I use the classification output as input for classification?
-You can but should be aware that iterative searches are error prone. We suggest to copy the folder into another folder (any name will do) and name the folder Training_Output. Then you need to generate a sequence file with all the sequences that have been classified so far and name this file TrSeqFile. That you need to put inside Training_Output. Should work.

3 Why is there no final colored tree with the classification?
-Because that would require the complete tree to begin with in the first place.

4 My overall classification is quite well but certain groups detected only few sequences?
There can be many reasons. In the original paper we show an example of what happens when the training is incorrect, in addition we show HMMERCTTER does not do well with sequences that contain repetitions. Another reason is that your target and training sets do not correspond. If a certain cluster is not well represented in a sequence set, few sequences will be accepted. On the other hand, the training set can lack sequences of a certain subfamily. HMMERCTTER cannot deal well with biased datasets.

5 Is a large training set by definition good?
-No. And this can be for several reasons. Certain superfamilies are too complex, or contain groups that cannot easily be discriminated by means of HMMERCTTER.

6 Things go wrong in the interactive phase of the qualification.
Sometimes the visual aids are not refreshed quickly enough, or sometimes you might have clicked on a button while HMMERCTTER was still busy. The latter tends to happen after the question if you want to accept additional positives. If you did accept additional positives, HMMERCTTER still needs to incorporate these and possibly more sequences, accepting positives is iterative. Be patient. To check: In the main MATLAB screen, it should state twice (group and complete target set) that a *hmmsearch* has been conducted in order to have finished the iteration.

7 *Error executing Dendroscope on scriptMidPointTree - /bin/bash: tcsh: command not found,*
You need to install tcsh and csh.

sudo apt-get install csh
sudo apt-get install tcsh

Bug reports

HMMERCTTER can be considered as a complex tool since it has many dependencies. Therefore we envisage things can go wrong. What to do when things go wrong?

1 Please check if the dependencies are functioning properly
2 Check the FAQ, maybe your problem has been described and solved.
3 Send us a report using the following guidelines.

> Describe the problem in the mail, indicating whether it concerns training or classification.
> Attach the output folder as a zipfile.
> Indicate in your mail details regarding your OS and the versions of the dependencies.

Please send mail to tenhave.arjen@gmail.com with "Bug report HMMERCTTER" in the subject and we will try to solve it as soon a possible.