Sequence bioinformatics, computer exercise XI
University of Potsdam, WS 2023/2024, Stefanie Hartmann

Give it all you've got! And since the mere presence of smartphones reduces available cognitive capacity (`https://www.journals.uchicago.edu/doi/full/10.1086/691462`), I suggest you put your cell phone out of sight and into your bag for the duration of this computer lab.

# Profile HMMs

# Objective

The goal of this exercise is to get more familiar with profile HMMs in the context of gene family analysis, and also to be able to use the popular software HMMer.

# 1 Locally installed software you'll need

To view last week's multiple sequence alignment again, you'll need an alignment viewer, e.g., JalView.

# 2 Logging into the cluster and preparing for the analyses

☞ Log into the university's computing cluster as described before and change to your directory for this class on the /work partition.

☞ You'll use the HMMer software today. It was written by Sean Eddy and implements profile HMMs for biological (protein) sequence analysis. The software and a very detailed manual is available at `http://hmmer.org/`. Load the HMMer module with the command
`module load bio/HMMER/3.3.2-gompi-2020b`

☞ Start an interactive session: `srun --cpus-per-task=1 --mem=5G -p teaching --pty bash`
Please coordinate with your neighbors so not everyone starts the interactive session at the same time.

# 3 The data

Today you will continue to work with the sequences from previous weeks. You will need for today's exercise:

- the sequence alignment from last week that you consider to be the best one
- the profileHMM file "PF01103.hmm", available in the shared folder on the cluster
- the file "moreProteins.fasta" from lab09 (BLAST) Note: the file can stay in the lab09 folder for today's exercise.

☞ Create a new directory for today's exercise and copy the files you'll need over.

☞ Recall or re-compute: Which species are represented in the file "moreProteins.fasta"? How many sequences per species are there?

# 4 Preparing: Seed alignments and profileHMMs

## 4.1 Obtaining a seed alignment

☞ A profile HMM is only as good as the alignment on which it is based, so take a good look at the MSA from last week that you consider to be the best one. Open the file with Jalview (or any other alignment viewer), and revisit the following question from last week: Do you think all of these sequences contain the same functional domain(s)? Are they globally homologous? And, therefore, can and should this set of sequences be globally aligned? yes, but not globally homologous. but they obiously have local alignment which is similar the front part can nat be homologues fuctional domain

☞ Furthermore, and very important for today's computer lab: can and should the entire MSA from last week be used to statistical model the different alignment regions? What are the assumptions and consequences of computing a profileHMM from a complete alignment or from a specific alignment region?

만약 전체를 대상으로 했을때는 낮은 퀄리티의 앞부분을 포함할 결과가 좋지 않을

not the entire regions

If you decided that not the entire alignment is suited as a seed alignment, how do you then decide which alignment columns and sequences to exclude? This is not an easy question to answer! **Below are some thoughts on this: carefully read this but don't start trimming your alignment quite yet.**

아마 가장 큰 chunk의 insertion이나 deletion을 가진 시퀀스가 제거될 수 있을 것.

**How to exclude or select sequences** Last week I already asked if some sequences are poorly aligned and should not be in the alignment. If you believe that a sequence is not homologous to the other sequences (or too divergent to be reliably aligned using a MSA software), it should be excluded.

☞ What might be the effect on the HMM of including a problematic sequence? of removing it? of removing too many sequences? of removing divergent but true members of the gene family?

1,2 . can reduce accuracy
3.

**How to exclude or select alignment columns** Without some experience with MSAs and gene family analysis, I don't recommend doing this based on your own judgement. Instead, you could use as a guideline the M-Coffee result and take only the (continuous!!) region that received high consistency scores. Alternatively, you could search one or more of the sequences against a domain database (e.g., Pfam) to see if and where there are conserved domains, then use the alignment regions that correspond to the functional domain.

☞ What might be the effect on the HMM of including alignment regions that violate positional homology?

☞ What might be the effect on the HMM of including alignment regions that contain more than one functional domain?

## 4.2 The OEP85 domain

The toc75 homologs you have worked with all contain a domain called OEP85. The OEP85 domain is part of eubacterial and eukaryotic proteins that function in protein translocation and in the assembly of membrane proteins. In order to carry out their function, many of these proteins also contain additional domains.

☞ Direct your internet browser to the Pfam entry of the OEP85 domain at `https://www.ebi.ac.uk/interpro/entry/pfam/PF01103/`. Briefly check out the information provided there, but don't spend too much time on this.

☞ Next to the name/ID of the Pfam Family, you'll see the number of architectures in which the domain is found. Click on this link: You'll see that the OEP85 domain is present in proteins with different architectures, and that these mostly differ in the domains present before the OEP85 domain. This might explain why the beginning of the sequences was difficult to align reliably, and why the beginning of the multiple sequence alignments you computed looked rather awful: it contains different and non-homologous regions! (Although I didn't investigate which of these domain architectures are actually present in our data set.)

For this exercise, you will (compute and/or) take a profile HMM only for the OEP85 domain of the proteins.

## 4.3 OPTIONAL: Computing a profile HMM

Here I'm describing how to compute your own profile HMM, although I don't recommend doing this during the 90 minutes we have for the computer lab. In the interest of time, please do this on your own. Skip to the next section now but feel free to come back to this section afterwards.

☞ Identify the alignment region that corresponds to the OEP85 domain. For example, search the TC753_ARATH sequence against the Pfam database online. Identify the position at which the OEP85 domain begins (ends) and delete all alignment columns that come before (after) this position:

> In JalView, click on an alignment position above the sequences, it will then be colored red. In the menu **Edit**, you can select **Remove Left** to delete larger blocks of the alignment to the left of that column.

☞ Ideally, you'd only use sequences for which you know that they are members of the gene family, for example based on external information. Without this information available here, use your own judgement. If you do want to remove one or more sequences, the new alignment might contain gap-only columns. In JalView, remove these via the menu **Edit** $\longrightarrow$ **Remove Empty Columns**.

☞ Save the new alignment under a new name in fasta format (menu **Edit** $\longrightarrow$ **Save As...**). Upload it to an appropriate location on the cluster using the `scp` command.

☞ On the computing cluster (with the HMMer module loaded and in an active srun session), compute the profile HMM using the syntax `hmmbuild <options> seed.hmm alignment.fasta` , substituting the names for input and output files appropriately. The default format for input alignments is the Stockholm format. To input a fasta file, use the `--informat afa` option. (A list of options is available using `hmmbuild -h`)

## 4.4 Using a profile HMM from Pfam

For now, please use the profile of the OEP85 domain that is available for download from the Pfam site (which I've already copied to the cluster). On the Pfam entry of the OEP85 domain, click on "Curation & model" (left hand menu). You'll find information about the model, including the number of seed sequences that went into generating it, and on the bottom you'll see a link for downloading the profileHMM. This is the one I've placed in the shared folder on the cluster, copy it to your own folder if you haven't done so already.

## 4.5 Understanding a profile HMM

The hmm file contains the profile Hidden Markov Model for sequences in the seed alignment from Pfam (or in your trimmed sequence alignment). Open it with `less`, with or without the "-S" option, depending on the width of your terminal.

The header section contains information about ID, number of sequences in the seed alignment, length of the model, and other details about the profileHMM. The line that starts with the tag 'LENG', for example, is the number of match states in this particular model.

The model parameters for the profile HMM begin on the line that starts with the tag 'HMM':

- it contains the symbol alphabet (the 20 amino acids) in the first line and the column headers for the state transition probability fields in the second line.
- the next three lines contain information about background frequencies and the BEGIN node. You can safely ignore them. The line that starts with a "1" is where the main model begins:
- for each match state (node), there are three lines with model parameters:
  - match emission line: this contains the match emission scores for each of the 20 amino acids for a given match state. (It also contains, at the very end, information about the MSA column, the consensus residue, and other annotation information of this alignment column.)
  - insert emission line: this contains the emission scores for each of the 20 amino acids for a given match state.

– state transition line: this contains the state transition scores for the transitions given in the line below the one that starts with "HMM".

Note that probability parameters are stored as negative natural log probabilities (e.g., a probability of 0.25 is stored as -log 0.25 = 1.38629. The special case of a zero probability is stored as '*'.)

☞ For the third match state of the profileHMM and for only the amino acids A, F, and L, draw the (properly labeled!) Hidden Markov Model.

## 4.6 hmmsearch

You will next use the profile HMM to search through the sequences in the file "moreProteins.fasta" for matches to this profile; note that you will not have to format the sequences for use as a database like you did for the BLAST search. As usual, options of the program can be listed with with `hmmsearch -h`. Your command will look something like the one below, but you will have to adjust it according to your file names (and to the path to the database if it's still in the folder for the BLAST lab):

```
hmmsearch seed.hmm moreProteins.fasta > moreProteins.hmmsearch
```

☞ Execute the hmmsearch-command and take a look at the resulting output file. It contains several sections in the style of a BLAST search result. Much more details about the output of hmmsearch, and how it is computed and displayed, is given in the HMMer manual. A brief overview of the most important parts are given here:

**header** this tells you what program was run, on which file, and which options you specified

**sequence top hits** this is a list of ranked top hits sorted by E-value, as in BLAST results. For each hit, the name of the target sequence, a description line for the sequence, the raw score (in units of 'bits'), the estimated E-value, and the total number of domains detected in the sequence are reported.
By default, every sequence with an E-value less than 10.0 is listed in this output. E-values of 0.01 or less are generally significant hits (above the 'inclusion thresholds'), and hits with E-values between 0.01 and 10 should be inspected manually.

- first three columns: E-value, score, (compositional) bias for the entire sequence. For single-domain hits, these values are (almost) identical to the next three values. For multi-domain hits, they represent the sum of all identified domains. This E-value is the one you should pay most attention to.
- next three columns: E-value, score, bias for the single best-scoring domain in the sequence. From the HMMer manual: "The idea is that we might be able to detect that a sequence is a member of a multidomain family because it contains multiple weakly-scoring domains, even if no single domain is solidly significant on its own".
- next two columns: expected and observed number of domains the target sequence contains
- last two columns: names and descriptions (if available) of the target sequences

**domain annotation for each sequence** is the next section. It contains for each sequence:

- a table of domain hits. It shows the domain matches within a given sequence and additional information about significance and position of each domain hit. The most important part for now is the independent E-value (*i-Evalue*): if it is significant, "even this single domain by itself is such a strong hit that it suffices to identify the sequence as a significant homolog".
- alignments in BLAST-style for each domain hit.
  - each domain alignments begins with a header that starts with "=="
  - the next line contains consensus annotation information for the column; in this example, you'll see a protein secondary structure consensus (https://tinyurl.com/ffm65se8).
  - the next line corresponds to the HMM consensus of the query model, where capital letters correspond to highly conserved residues. Dots indicated insertions in the target sequence with respect to the model.

  - the center line shows where the hit matches the HMM, and where there is a positive score ("+"), or a "conservative substitution" with respect to what the model expects at that position.
  - next is the target sequence. Gaps indicate deletions in the target sequence with respect to the model
  - the last line represents the expected accuracy of each aligned residue; values range from 0 (0-5%) to * (95-100%) and represent posterior probabilities computed by HMMER3.

**'the rest'** contains some summary statistics.

☞ Based on the result from `hmmsearch`, how many of the sequences in the database file would you consider 'significant hits' to the HMM? Which would you exclude, and why? By looking at the alignments, decide on an E-value threshold for hits you consider to be significant. 한 30개쯤 되는듯

☞ Repeat the hmmsearch, but this time use the E-value threshold you decided on. The option for this is `-E` (one dash!), followed by the value. Note that the options need to be specified before the .hmm and .fasta files. Take a look at the new output and verify by inspecting the alignments that all reported sequences are indeed significant hits. If necessary, adjust the E-value threshold again.

☞ Run `hmmsearch` with the option `--tblout` (two dashes!) and followed with a file name for the table; this will generate the output in table format with one line per sequence hit. Execute this search.

☞ Extract the database sequence IDs from this table using system utilities. Compare with the results from the BLAST search (that was done against the same database!).

## 4.7 Aligning hits to the profileHMM

For further analysis of this gene family, you might want to compute an alignment of all hits from the hmmsearch. There are several options for doing this. You'll only use the second of the following options, but make sure you understand the difference between them, and their potential applications, advantages, or disadvantages:

A. extract the hit sequences by their IDs from the database file and align them using any of the MSA programs you used last week. This would result in a *de novo* alignment of the full length sequences.
B. use the option "-A" (followed by a file name) for the `hmmsearch` program to compute and save an alignment of the hits to the profileHMM. This would result in an alignment that covers only the region of the profileHMM, and the beginning and ends of the hit sequences would **not** be included at all.

☞ Execute the `hmmsearch` program once again, with the E-value cutoff you determined before. This time also output an alignment as described above, by including the option "-A", followed by a file name for the alignment: Save the resulting alignment in a file with the extension ".sto", the recommended extension for the Stockholm format.

☞ Take a look at the resulting alignment using `less -S`. The alignment is in HMMer's native format, the Stockholm format (https://en.wikipedia.org/wiki/Stockholm_format), which is also used by the Pfam and Rfam databases:

- the format allows extensive annotation (mark-up) of the alignment: Information about the alignment "(`#=GF`)", about a sequence "(`#=GS`)", or about an alignment column "(`#=GC`)" can be included. Here, you'll see the expected accuracy per residue for its alignment to the profile HMM.
- the '-' and '.' characters both represent gap characters: The '-' indicates that the sequence is missing a column that the HMM was expecting to be there. The '.' means that in this alignment position at least one sequence has an insertion, compared to the profile HMM.
- lower-case letters indicate residues that are emitted from the insert state.

This alignment might be easier to read and interpret when opened in an alignment viewer. JalView can read alignment files in stockholm format, so you can use `scp` as before and copy the alignment to your local

computer. When trying to open the file from within JalView, make sure the "Format" option is set to either "All Files" or to "Stockholm".

> Note: if the alignment viewer you use does not support the stockholm format, you can convert it to a fasta file. For, this, first load the EMBOSS module (`module load bio/EMBOSS/6.6.0-foss-2018b`), and then use its seqret program (e.g., `seqret -sequence aln.sto -outseq aln.fa -osformat fasta`) for the conversion.

☞ Compare the alignment to the *de novo* alignments computed last week.

☞ Are you impressed by the quality of the alignment obtained in such a short time?

☞ Considering what you learned in lecture about profile HMMs, and considering previous exercises, this exercise, and the final alignment: list at least two ways for improving the final alignment.

# 5 Exiting the terminal

Log out of the interactive session with `exit`.

Then log out of the current session to the cluster, type `exit` again.

Finally, to cleanly close the terminal window, type `exit` yet again.

Do not close a terminal window with the mouse, especially not if an ssh session is still active.

## ➜ Don't forget to take your flash drive with you when you leave!