Project CL0057 summary
Contract: 388/313-14-00/DP/2020

Maciej Sikora, Vasilina Zayats

Spis treści

1 Clan analysis

1.1 Clan description

Met_repress (CL0057) is a superfamily that contains many antitoxin families, all of which carry a ribbon-helix-helix DNA-binding motif with the beta-ribbon located in and recognizing the major groove of operator DNA.

1.2 Preprocessing

Families in clan CL0057 can be separated into 3 groups:

- Short (one-domain) unfused unknotted families.
- Long (two-domain) fused unknotted families.
- Long (two-domain) fused knotted families.

Sequences preprocessing steps:

- Sequences from our clan were downloaded from Pfam, filtered by length, and by similarity, using Cd-Hit 90%.
- Families: PF07181, PF08870, and PF10802 were also separated into domains (two for each).
- Families PF07181 and PF10802 were enriched by additional sequences from Uniprot (and separated into two domains).
- Additionally, we added one more unknotted fusion sequence from PDB structure 4hv0. (Described as PF00000) (Discussion?)

1.3 CLANS

The problem here was that many Pfam families didn't cluster into families, and instead, produced "cloud". There are no clear borders between many families. On the other hand, there were also families located quite far from others. To solve "cloud" problems, we tried two solutions. Alternative clustering using MCL. Because separating sequences based on families failed, the idea was to create alternative clusters based on sequence similarity (BLAST). We tried various parameter values from 1.1 (too large clusters) to 6 (massive amount of small clusters (1-2 sequences)) with the best results using values between 1.4-1.5. Unfortunately, while MCL presented slightly better clusters, even then sequences were mixed between clusters. Similarly, we tried using several different methods - both sequence and statistically based without success. All methods were classifying the main cloud as one-two clusters or many groups without clear borders. Instead, we tried clearing sequences even further while using standard family-based separation. So we used Cd-Hit again, this time with lower percentage values. We were using one or two-step filtering. The first part was to run Cd-Hit 50% clusters, and instead of taking output sequences (such filtering would be too strict), we deleted clusters with 1–2 sequences only. If there were still problems with clustering, we were rerunning Cd-Hit 70% and again deleted only small clusters with up to 3–4 sequences. Such a process is reducing the outer part of the CLANS clusters - sequences mostly invalid or different from the rest while leaving core, family representative sequences mostly untouched. This approach gave much better results - the "cloud" was still there, but much cleaner.

1.3.1 Cloud problem

- MCL clustering - no desired effects.
  - Parameters tested from 1.1 to 6.
  - Best values - around 1.4-1.5 - Still clusters were mixed and didn't solve
- Reducing the number of sequences using Cd-Hit with a lower percentage: 50–80% (Depending on the family). This partially solved the problem as the cloud was slightly more clearer, however, it didn't completely separate families.

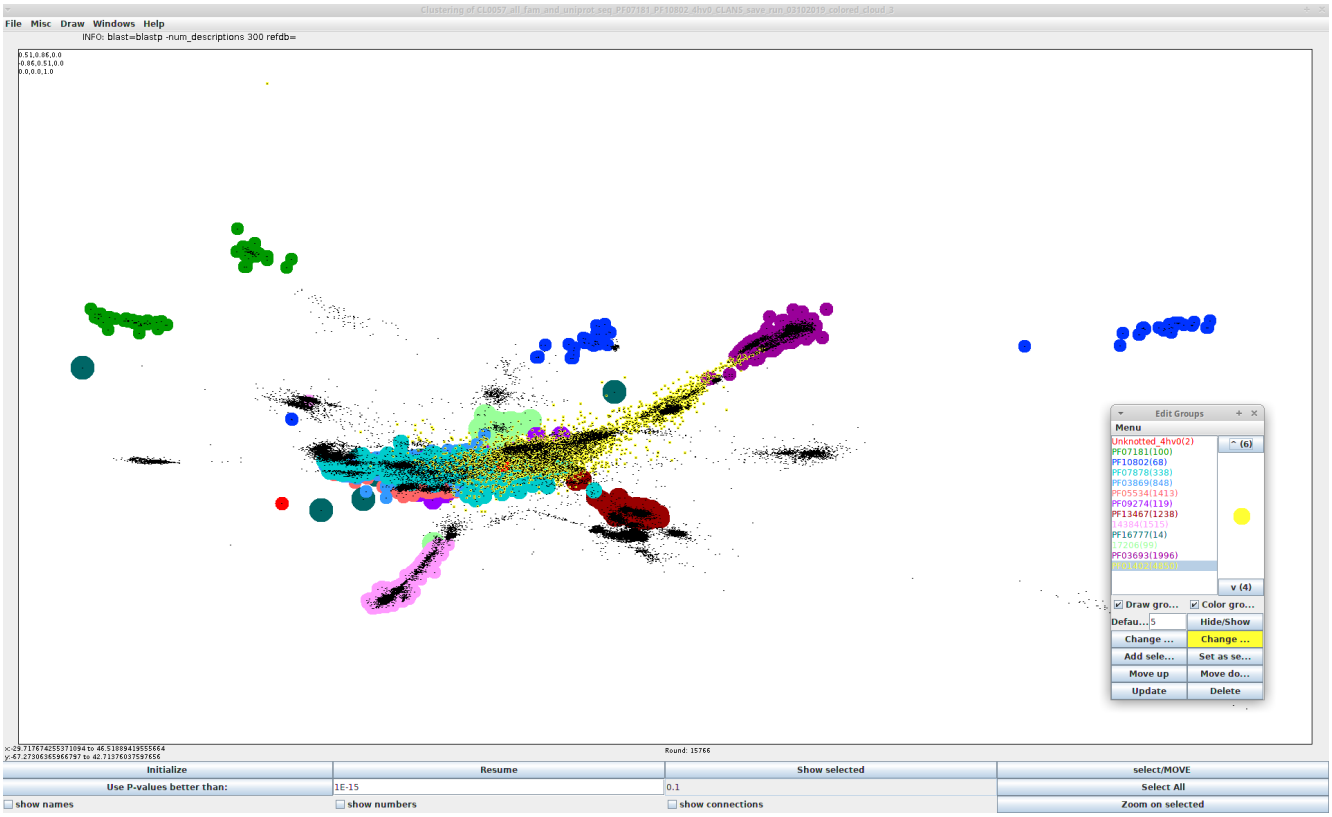


Fig. 1: CL0057 clustering in CLANS. Family PF08870 is missing (∼300 seq).

1.4 Ola's Workflow, profiles

Two families were ignored by workflow because of their e-value bigger than 0.01. Another test shows that those two families were also much different from each other.

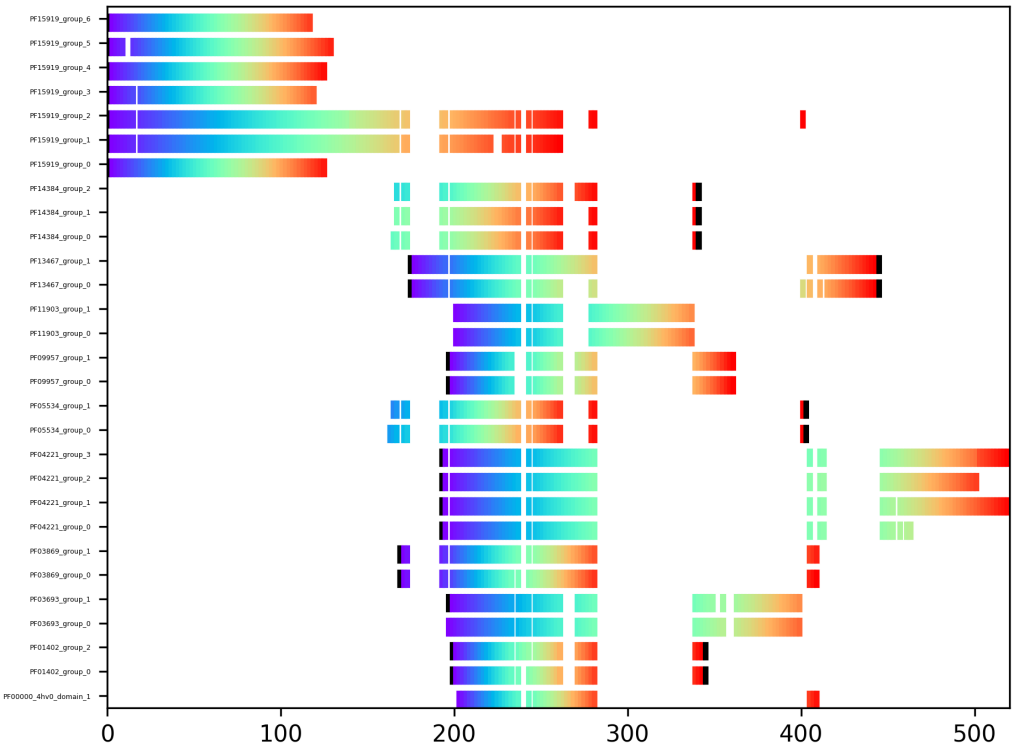


Fig. 2: Plot from Ola's Workflow with families from clan CL0057.

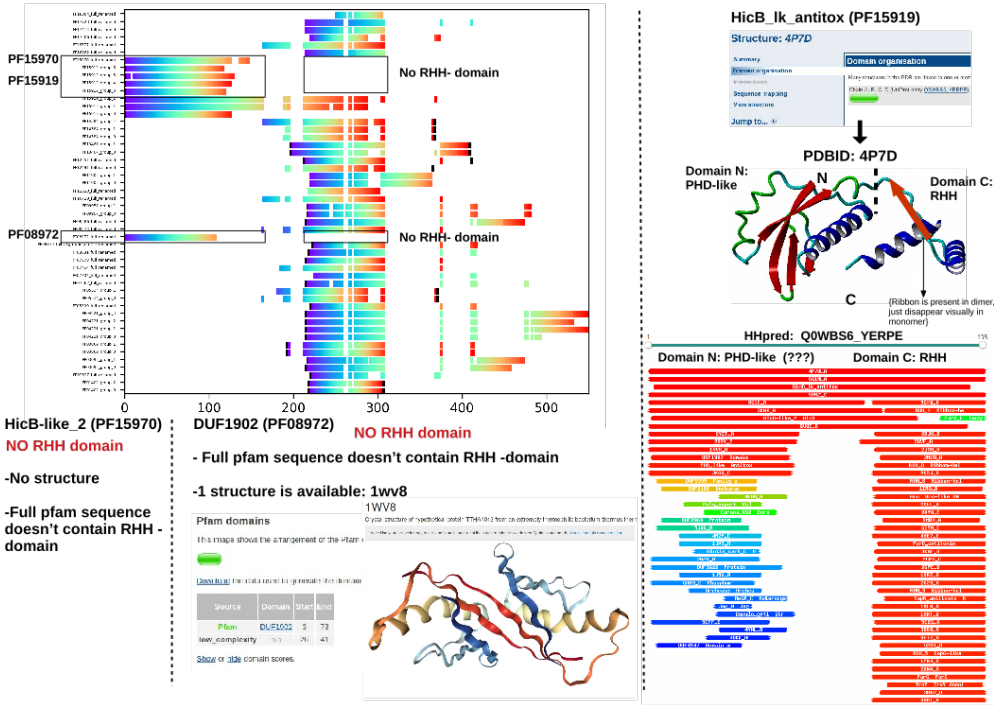


Fig. 3: Clan CL0057 domain analysis.

1.5 Clan trees

I decided to start by creating a tree with families from our clan only to tune parameters. There was some progress from "brush" only trees to much better-looking ones. However, their quality was far from perfect, so I consider this part as not finished yet.

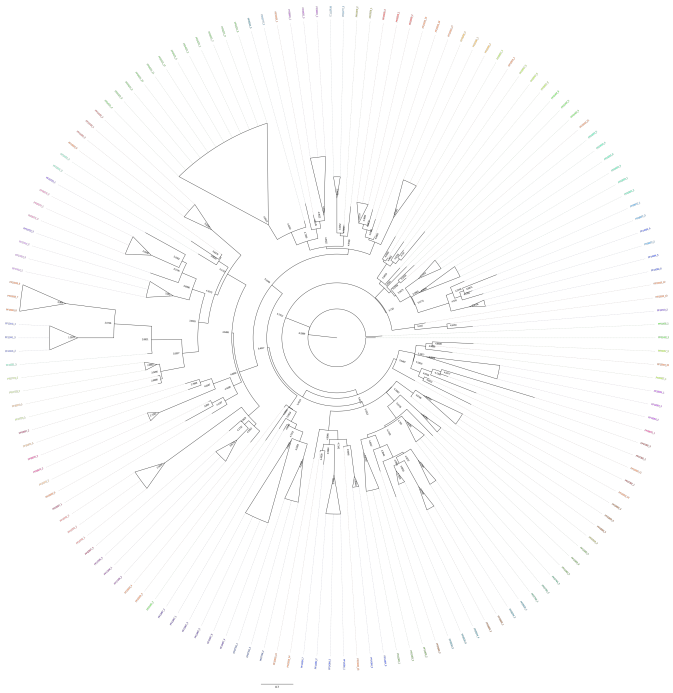


Fig. 4: Best old tree.

1.5.1 Automatic model selection tools

- SMS: Both AIC and BIC criterium are suggesting Vt +G+F model.
- IQTREE model finder - VT
- MrBayes - Blosum

1.5.2 First tree

Preprocessing:

- Families listed above were downloaded from Pfam (domains, full, unaligned)
- Reduced using Cd-Hit 80%
- Renamed using script.

1.5.3 Best tested trees

In the results below tree calculated using 4 chains reached 0.016 SD and using 12 chains - 0.022 (goal is to reach 0.01 or lower, while range 0.01-0.05 is acceptable). PSRF (Possible Scale Reduction

Factor) values were around 0.9998 (So very close to target: 1) However, those values probably can still be improved after more generations.

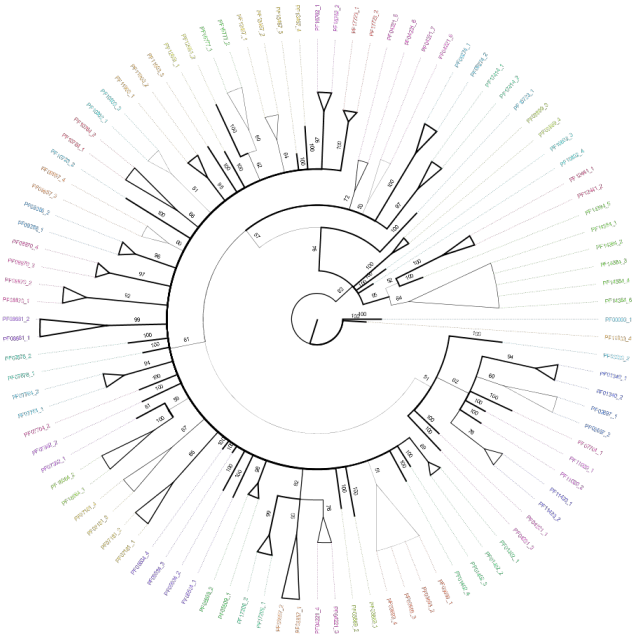


Fig. 5: 4 chains long run tree.

Fig. 6: 12 chains short run tree.

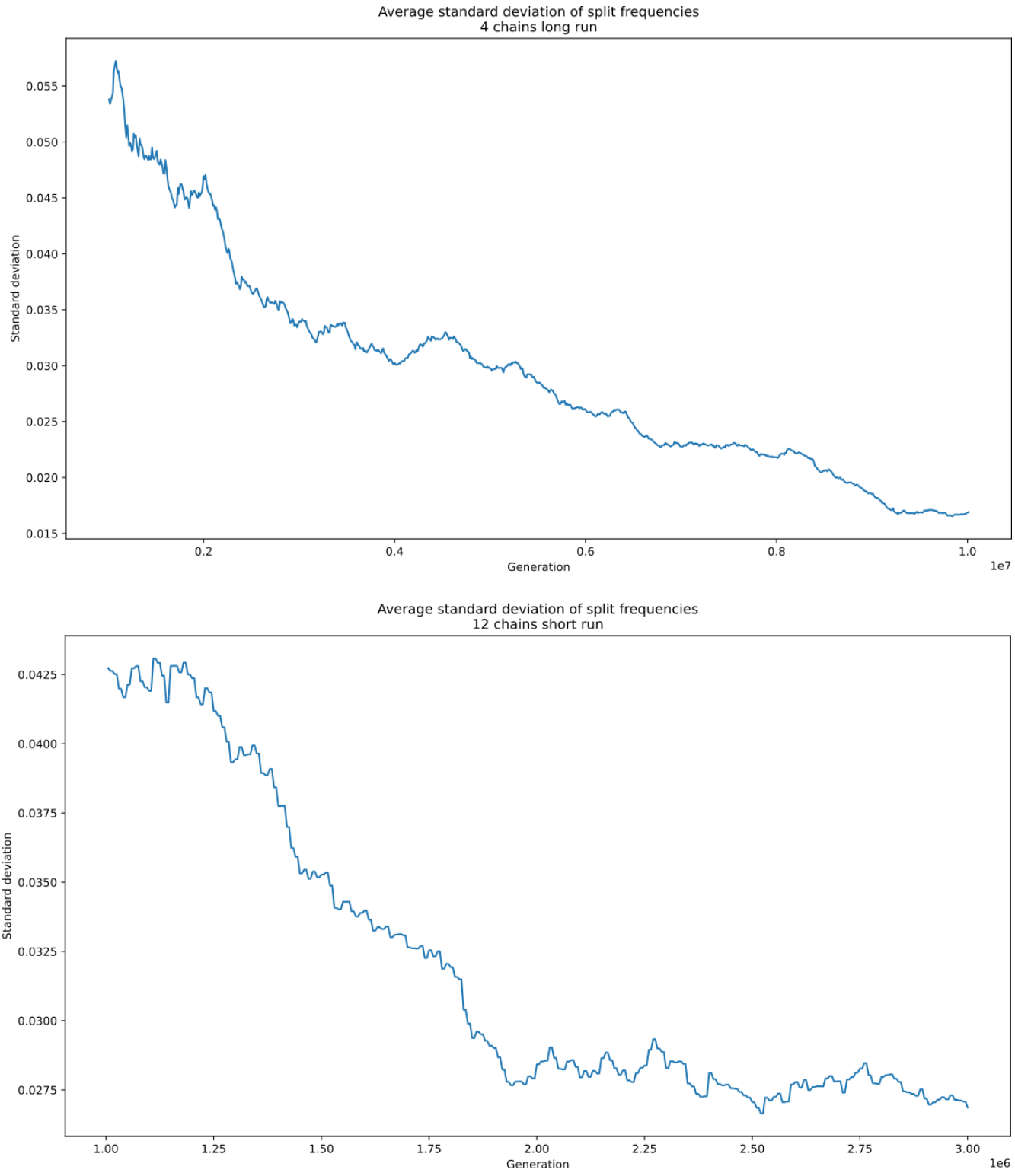


Fig. 7: The average standard deviation of split frequencies during both runs. Trees are trying to reach 0.01 values, however, 0.01-0.05 is still acceptable. Plots are starting from 1.000.000 generation. Notice, that 4 chains run has 10.000.000 generations while 12 chains run only 3.000.000.

2 New outgroups checklist

2.1 Looking for outgroups, problems

There are 37 families in our clan. Such a number makes analysis much harder to perform - a combined file is too big for most services or programs. Separately analyzing each family was very time-consuming and carried the risk of errors. This encouraged us to create a script that makes the process of searching on JackHMMER automatic. Results are sorted, pre-analyzed, and available offline with direct links to the website. With ready summaries, outgroups were picked and analyzed manually.

2.2 CL0123

HTH (CL0123) could be an outgroup. It is a large clan of DNA-binding domains that contain a helix-turn-helix motif. Several families from CL0123 were found by JackHmmer as possible outgroups for CL0057: HTH_3 (PF01381), HTH_26 (PF13443) and UPF0175 (PF03683). Although some of these motifs contain ribbon, the ribbon doesn't superimpose to the ribbon of Met_repress (RHH). However, two helices can be superimposed to helices of RHH .

Two helical regions of HTH superimposes to two helices of RHH-motif. Also, HTH_3 and HTH_26 domains contain two copies of the helix-turn-helix motif (2xHTH). However, two copies do not look the same (do not superimpose perfectly), perhaps they already adjusted to function as one domain and adapted their structures. RHH (two helices of RHH_1 fam) can be superimposed roughly to any of them, but still, it fits better to one of the two copies. Since there are both sequence and structure similarity found between these families, they (HTH families) can be considered as an outgroup for CL0057. However HTH CL0123 is a large clan, so probably it is better to select several families from the whole clan.

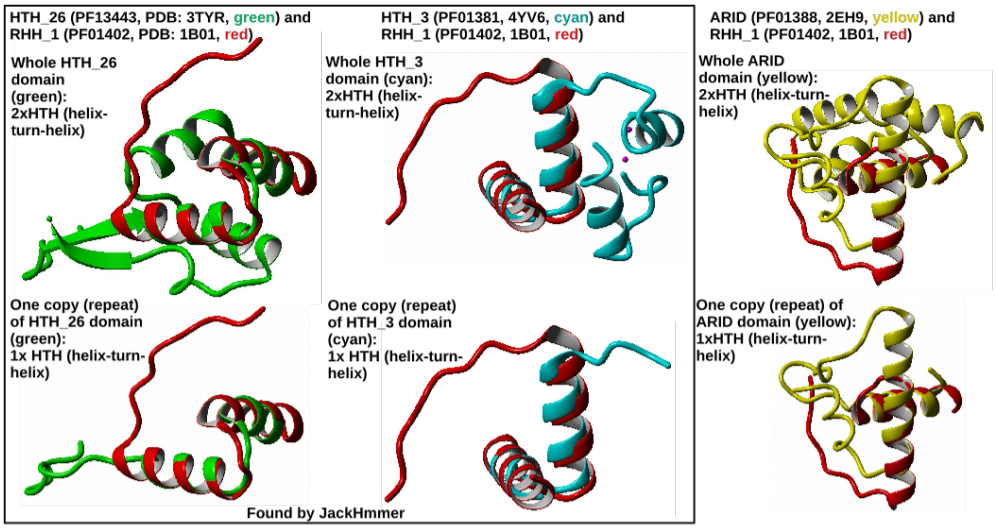
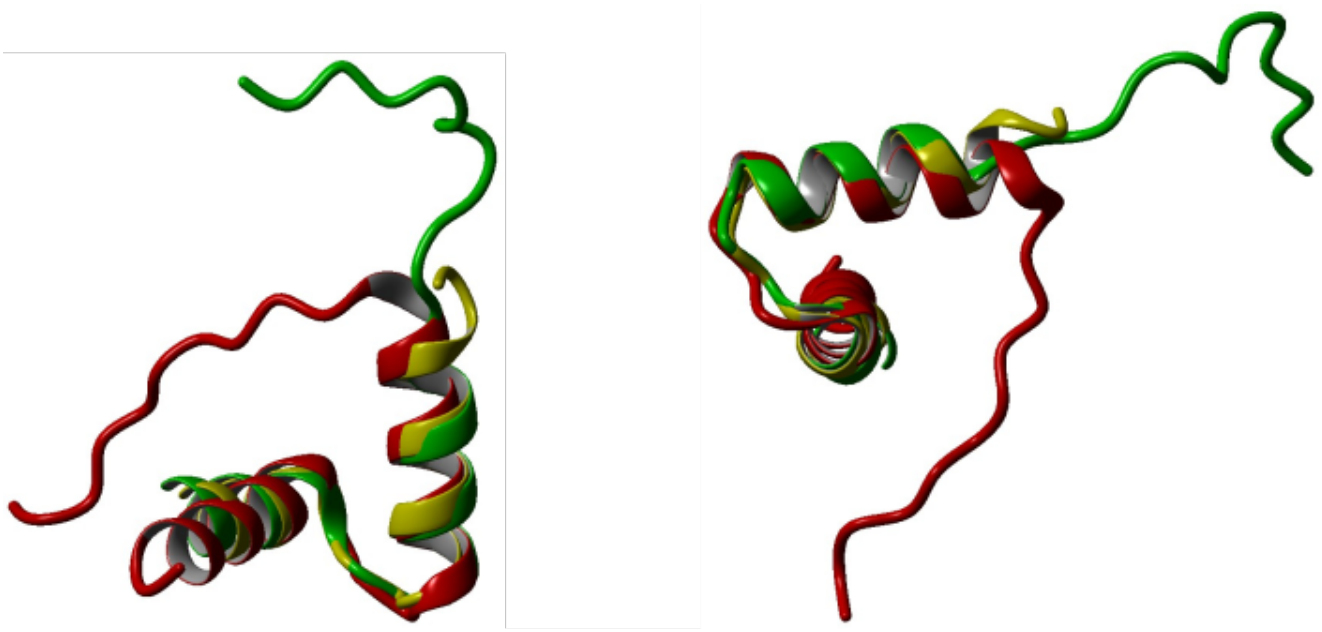
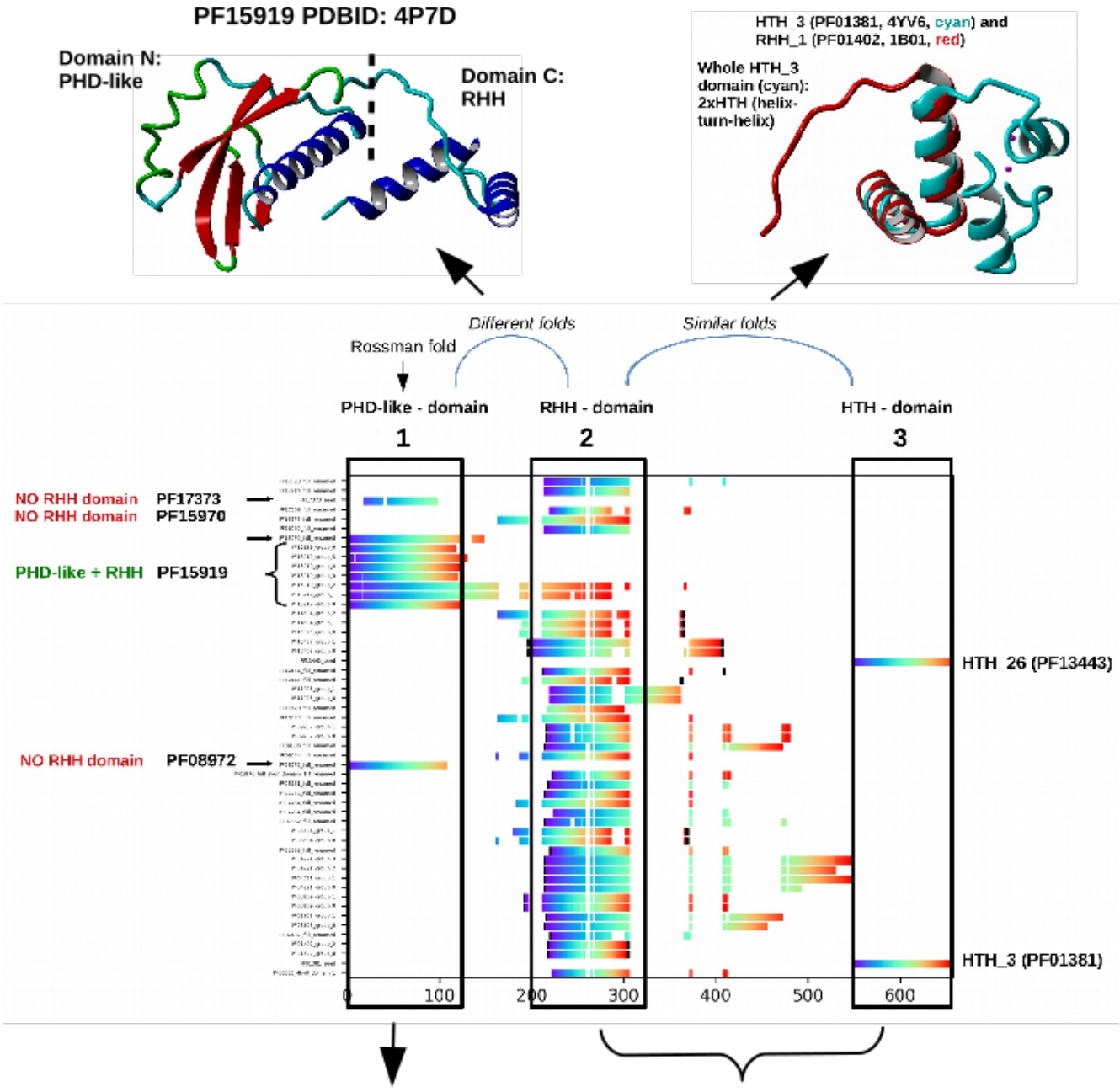


Fig. 8: Superimposition of RHH_1 domain (Met_repress CL0057) to HTH domains (HTH CL0123)



```
Conservation:                                                99
2wvb_obj_9                      1  -DSIIRFSVSLQQNLLDELDNRIIKNGYSSRSELVRDMIREKLVE   44
2wvb_1mon.pdb_chainA_s003       1  -DSIIRFSVSLQQNLLDELDNRIIKNGYSSRSELVRDMIREKLVE   44
3qq6A_obj_8                     1  ---HHMI--------GQRIKQYRKEKGYS-LSEL-AEKAG-----   27
3qq6_1mon.pdb_chainA_s002       1  ---HHMI--------GQRIKQYRKEKGYS-LSEL-AEKAG-----   27
3tyr_HTH_26A_obj_2              1  RENLYFQGMRVS---YNKLWKLLIDRDMK-KGEL-REAVG-----   35
3tyr_1mon.pdb_chainA_s001       1  RENLYFQGMRVS---YNKLWKLLIDRDMK-KGEL-REAVG-----   35
Consensus_aa:                      ...hh.b.........pcl.p..bcpshp..tEL..-.h......
Consensus_ss:                         eeeh        hhhhhhhhhh     hhhh hhh
```

Fig. 9: Superimposition and structural alignment (Promals3D) of RHH 1 structure (2WVB) and outgroups from clan CL0123 HTH 26 (3TYR), HTH 3 (3QQ6). Original file source: Outgroups_HTH_and_RHH.



This is a different domain – PHD-like – not similar to RHH domain. Therefore families which have only this domain can be removed from the analysis: PF17373, 15970 and PF08972. Also, it would be better to remove this domain (PHD-like) from the family PF15919.

These domains (RHH and HTH) have similar folds. Two helices superimpose well. However, sequences of HTH_26 and HTH_3 families didn't align to RHH - they are not significantly similar at sequence level.

2.3 Selected families and outgroups

After further analysis correct outgoups were chosen.

- PF05261 - Tra_M / DNA-binding - CL0548 (IHF-likeDNA-bdg)
- PF13443 - HTH_26 - CL0123 (HTH)

- PF01381 - HTH_3 - CL0123 (HTH)

From clan CL0057 only families with RHH were picked.
Those files were the base for further analysis.

3 Combined analysis

3.1 Ola's Workflow, profiles

Out of those eight outgroups (PF05261 and PF01388 to check again probably), only three passed the 0.01 e-value threshold in Ola's workflow, and only one aligned with the secondary column.

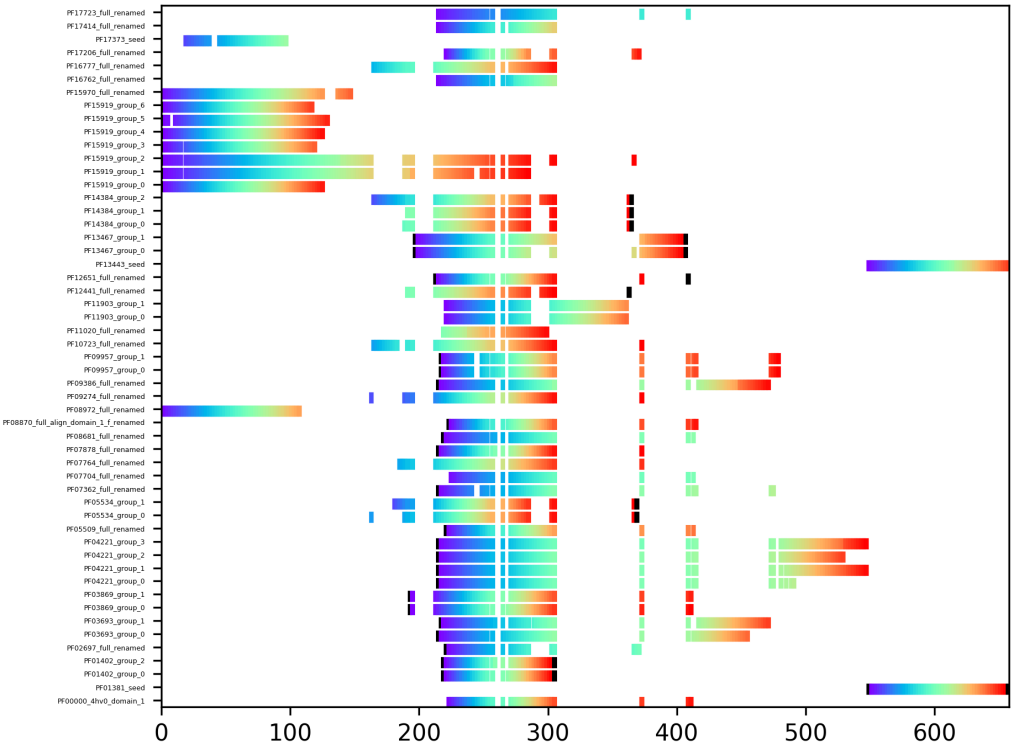


Fig. 10: A plot from Ola's Workflow with families from clan CL0057 and outgroups.

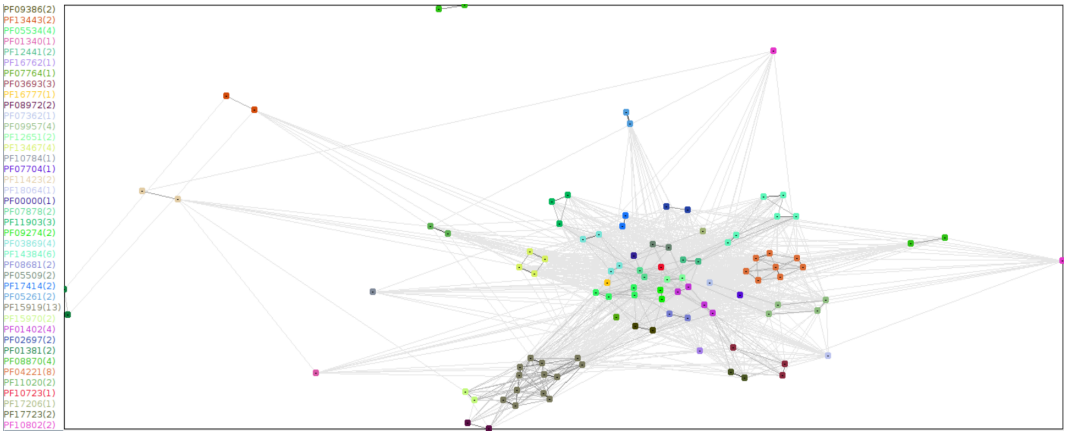


Fig. 11: Porfiles clustering in CLANS.

4 Used scripts

4.1 Trees part

To speed up trees coloring after calculating, I wrote a new script. As input, it takes the best tree from MrBayes and adds a new tag (like probability, length, etc.) to leaves – family. Each family has its own identifier, so every representative sequence from it will have the same id. This improved tree can be loaded in Figtree and using an internal tool, color by this tag.

To analyze the quality of the trees, I'm using another script that reads the standard deviation of split frequencies. This data is presented on a plot (examples below). Additionally, the plot can start from later generation (I'm using 1.000.000) to precisely show smaller values (Values before that point are unstable).

4.2 Clustering

To automate process of clustering low percentage Cd-Hit was used (psi-cd-hit). Here two additional scripts were needed. First is taking cd-hit .clstr file as input, and preparing files with groups. User can also specify minimum of seequences in cluster to get qualified. If this option is used, all small clusters are saved into one file.

Folder with group files and CLANS savefile are the input for the second script. Script is reading inforamtions from CLANS savefile and writes a group section into it. Sequences names need to exactly match those in savefile to be recognised. As the point of the script is to use the same input file for CLANS clustering and those 2 scripts, there should be no issues.

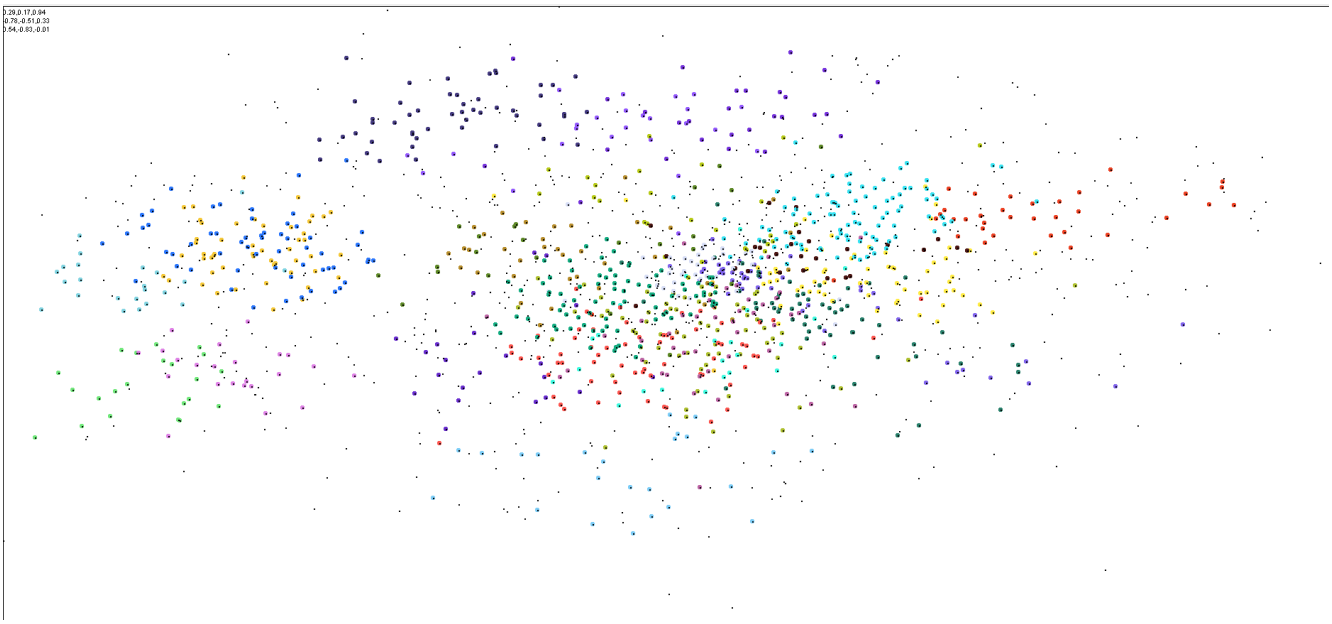


Fig. 12: Example result of using those 2 scripts. Clustering of family PF01388.

5 Used files

Sequences were downloaded from Pfam as fasta files, all the preparations were made on this extension as well. Later programs for tree creation required specific files versions. To convert between formats and visually check quality of the sequences program AliView was used. Files were converted into Phylip for PhyML and FastME programs and into nexus file for MrBayes. Result files were different from each tree building algorithm, yet all consensus trees were based on .json style type of the tree. Some algorithms were expanding this format with additional information like sequence names directly built in. Another expansion was the family id as a random and itself meaningless number which later helped with much easier tree coloring. All the consensus trees are possible to import into popular tree visualization programs like FigTree used in this project. Many additional files were created such as .mcmc statistic file from MrBayes, profiles or CLANS clustering files. All of those files are precisely described in manuals to those programs. No unusual modifications were made there other than analysis of those files or appending standard information via extra scripts (like coloring groups in CLANS).

6 All tested trees

- GORDIAN77
  drzewa_no_RHH
  lset nst=6 rates=invgamma, ngen=3.000.000
  10.000 - 0.34
  100.000 - 0.11
  1.000.000 - 0.06
  2.000.000 - 0.04
  3.000.000 - 0.033
  best results 0.033

- GORDIAN77
  drzewa_no_RHH_6_chains
  prset aamodel=mixed, Nchains=6 ngen=3.000.000
  10.000 - 0.33
  100.000 - 0.09
  1.000.000 - 0.05
  2.000.000 - 0.032
  3.000.000 - 0.0265
  best results 0.025929 (2975000) (continue?)

- GORDIAN77
  drzewa_no_RHH_12_chains
  prset aamodel=mixed, Nchains=12 ngen=3.000.000
  10.000 - 0.33
  100.000 - 0.094
  1.000.000 - 0.042
  2.000.000 - 0.028
  3.000.000 - 0.027
  best results - 0.0269

- GORDIAN77
  drzewa_no_RHH_8_chains_higher_nswaps
  prset aamodel=mixed, Nchains=8 ngen=1.000.000 nswaps=2
  10.000 - 0.33
  100.000 - 0.10
  1.000.000 -
  2.000.000 -
  3.000.000 -
  best results -

- GORDIAN77
  drzewa_no_RHH_8_chains_even_higher_nswaps
  prset aamodel=mixed, Nchains=8 ngen=1.000.000 nswaps=4
  10.000 - 0.345
  100.000 - 0.093
  1.000.000 -
  2.000.000 -
  3.000.000 -
  best results -

- GORDIAN77
  drzewa_no_RHH_8_chains_2_nswaps_higher_temp
  prset aamodel=mixed, Nchains=8 ngen=1.000.000 nswaps=2 temp=0.3
  10.000 - 0.342
  100.000 - 0.115
  1.000.000 -
  2.000.000 -

3.000.000 -
best results -

- GORDIAN77
  drzewa_no_RHH_8_chains_2_nswaps_lower_temp
  prset aamodel=mixed, Nchains=8 ngen=1.000.000 nswaps=2 temp=0.05
  10.000 - 0.343
  100.000 - 0.091
  1.000.000 - 0.0396
  2.000.000 - 0.033
  3.000.000 -
  best results -

- GORDIAN77
  drzewa_no_RHH_8_chains_2_nswaps_LG
  prset aamodel=fixed(LG), Nchains=8 ngen=1.000.000 nswaps=2
  10.000 - 0.341
  100.000 - 0.0957
  1.000.000 - 0.058
  2.000.000 - 0.041
  3.000.000 -
  best results -

- GORDIAN77
  no_RHH_8_chains_2_nswaps_SMS_params
  prset aamodel=fixed(vt) statefreqpr=fixed(empirical)
  lset rates=gamma, Nchains=8 ngen=1.000.000 nswaps=2
  10.000 - 0.33
  100.000 - 0.106
  1.000.000 -
  2.000.000 -
  3.000.000 -
  best results -

- GORDIAN77
  drzewa_no_RHH_8_chains_2_nswaps_VT
  prset aamodel=fixed(vt), Nchains=8 ngen=1.000.000 nswaps=2
  10.000 - 0.338
  100.000 - 0.0947
  1.000.000 - 0.052
  2.000.000 -
  3.000.000 -
  best results -

7 Old outgroup checklist

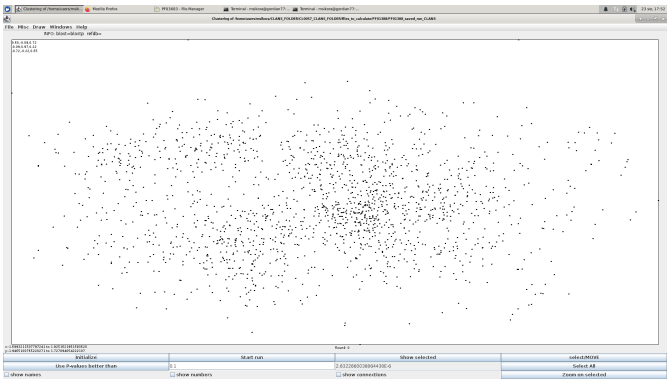
7.1 Other families

7.1.1 PF01388 - ARID - CL0123 (HTH)

Although ARID (PF01388) was not found by JackHmmer, there is a structural similarity between RHH_1 and ARID and likely to other families from the HTH clan.

- CLANS

Family is rather small and separation I think separating into groups shouldn't be considered here.

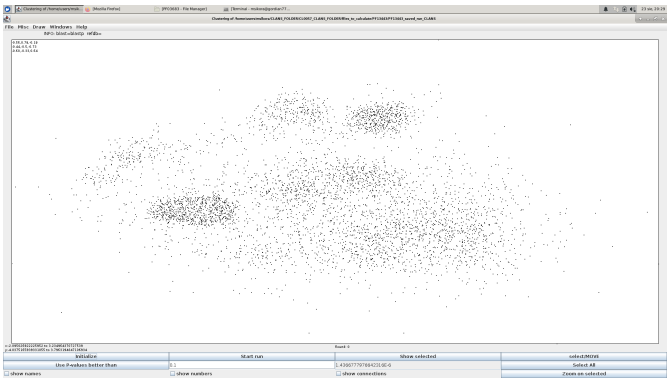


7.1.2 PF13443 - HTH_26 - CL0123 (HTH)

PF13443, PF01381 aligned with each other, because they are from the same clan. Although HTH fold is similar to RHH, there is no significant sequence similarity.

- CLANS

Not so clear groups, however, it might be possible to separate it into 5 groups (4 smaller and 1 big).

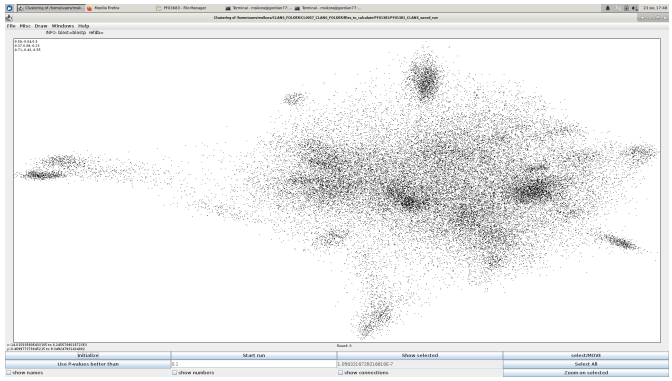


7.1.3 PF01381 - HTH_3 - CL0123 (HTH)

PF13443, PF01381 aligned with each other, because they are from the same clan. Although HTH fold is similar to RHH, there is no significant sequence similarity.

- CLANS

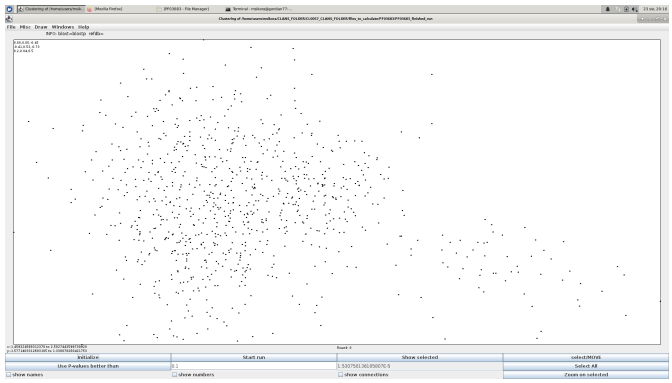
Separating into 1 main group and 4 subgroups (or just 1 extra) may be considered.



7.1.4 PF03683 - UPF0175 - CL0123 (HTH)

- CLANS

I wouldn't separate this family into groups, as one group wouldn't have enough sequences.

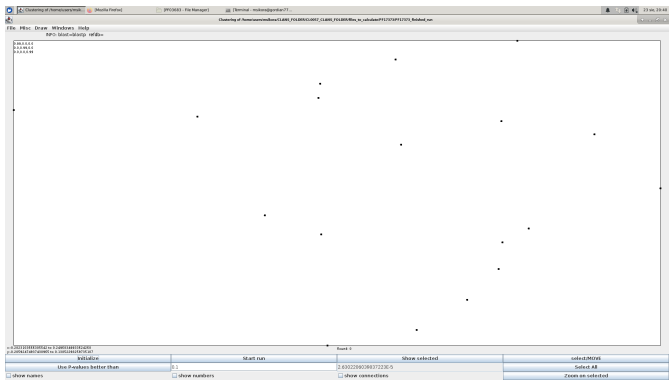


7.1.5 PF17373 - DUF5395 - No clan

Not aligned to the RHH column. Aligned to PHD-like domain of PF15919 (1st column) DOMAIN of PF15919 - HicB_like antitoxin of bacterial toxin-antitoxin system!!! (Difference between Pfam and . . . ?)

- CLANS

Not enough sequences for separation.

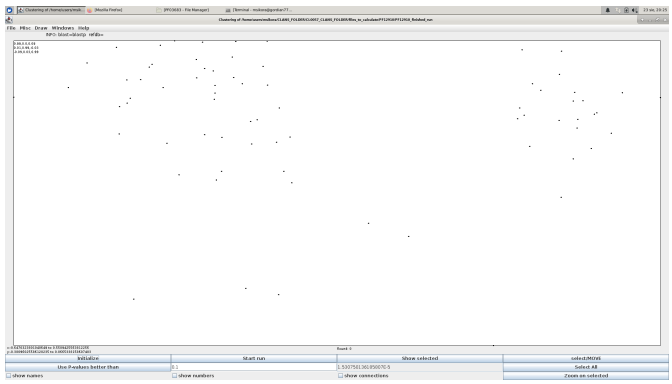


7.1.6 PF12910 - PHD_like - CL0136 (Plasmid antitox)

PHD_like (PF12910) - this domain appears to be the N-terminus of the RelB antitoxin of the toxin-antitoxin stability system. Therefore it is related in function to Met_repress. However, HicB_lk_antitox (PF15919) Pfam domain has a long seq (140 residues), so it contains not only the RHH motif but also another domain. PHD_like (PF12910) aligns with another domain of HicB_lk_antitox (not RHH motif), therefore it appears in this search, but it is not relevant for RHH motif.

- CLANS

Separating into 2 groups may be considered, however they will have only few sequences in each.

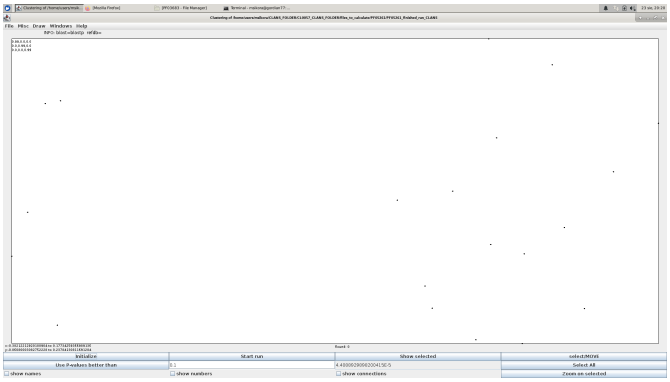


7.1.7 PF05261 - Tra_M / DNA-binding - CL0548 (IHF-likeDNA-bdg)

Aligning with center column.

- CLANS

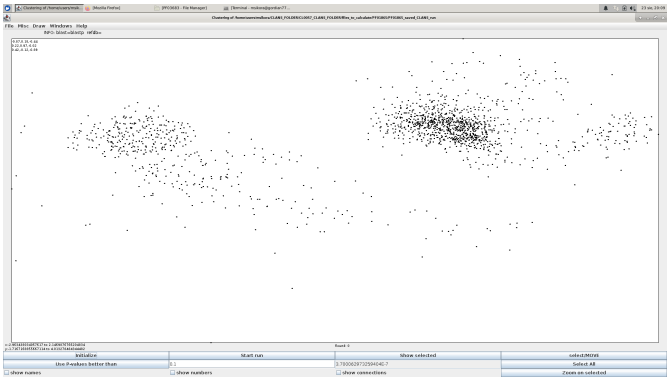
Nothing to separate - there are only 19 sequences in the family.



7.1.8 PF01865 - PhoU_div / DUF47 - CL0297 (PhoU)

- CLANS

Separation into 2 groups may be considered as seen below.



7.1.9 PF00096 - Zinc finger, C2H2 type - CL0361 (C2H2-zf)

Domains are very short (around 20 AA), and there are way too many sequences.
I was able to push the family into Ola's workflow, no results.

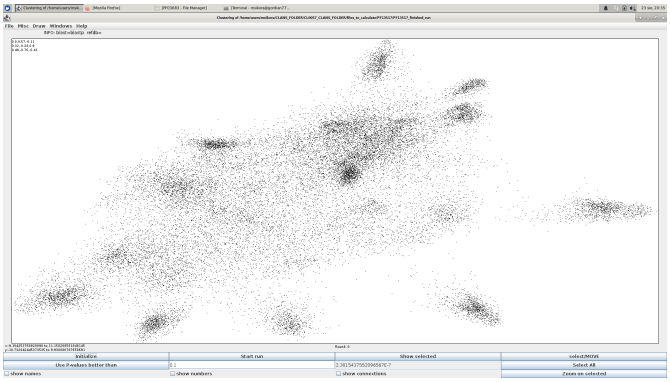
- CLANS

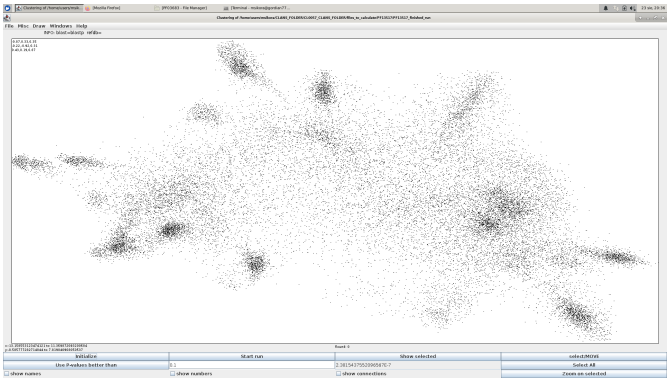
Unable to finish CLANS . First part - BLAST - finished, however, I'm unable to load sequences into second part - clustering - Out of memmory errors.

7.1.10 PF13517 - VCBS / Repeat domain in Vibrio, Colwellia, Bradyrhizobium and Shewanella - CL0186 (Beta propeller)

- CLANS

It might be possible to separate each "spike", but what about middle cloud? I'm not sure about this family, however it's definitely a big one, so separation may be suggested.





7.1.11 PF05792 - Candida_ALS / Candida agglutin-like (ALS) - No clan

- CLANS

Separating into 2/3 groups may be considered, but probably not necessary.