HMM search output analyses

LJM

2019-12-03

Contents

Chapter 1

Load libraries

```
# This is needed for spread()
library(tidyr)
library(tidyverse)
```

Chapter 2

Initilise variables

```
#fOTUsHMM <- read_csv(file = "../analyses/fOTU_HMM_headers.csv")
fOTUsHMM <- tibble()

path_to_data <-
  "../analyses/HMM_scan_using_eggNOG_HMMs/hmmsearch_out_parsed/"

files <- list.files(path = path_to_data,
                    pattern = "*.tsv", recursive=FALSE)

# Read in some data
tsv_names <- c("fOTU_name",
               "hmm_profile_id",
               "inside_inclusion_threshold",
               "Target_Bin_id",
               "Target_Seq_id",
               "full_sequence_e_value",
               "full_sequence_score",
               "full_sequence_bias",
               "best_one_domain_e-value",
               "best_one_domain_score",
               "best_one_domain_bias",
               "exp",
               "N",
               "description")
```

Chapter 3

Read data and create a tibble out of them

```
# Go through each tsv parsed results file
for(fOTU_file in files) {
  # Create the file path
  path_file <- paste(path_to_data,fOTU_file, sep = "")

  # Create a tibble from a tsv file
  fOTU <- read_tsv(file = path_file,
                   col_types = "cclcfddddddddic",
                   col_names = tsv_names) %>%
    # Take only interesting columns
    select(.,c("fOTU_name",
               "hmm_profile_id",
               "full_sequence_e_value")) %>%
    # Create a new col with -log_10 e-values
    mutate(log_eval = -1*log10(full_sequence_e_value)) %>%
    # Drop the old e-value column
    select(.,-full_sequence_e_value)

  # Check if there were empty bins
  fOTU_first_term <- fOTU %>%
    pull(fOTU_name) # This returns a vector of fOTU names
  # Pick just first element because that
  # would be where my "dummy" text would reside
  fOTU_first_term <- fOTU_first_term[1]

  # Check if the results had no hits
```

```
  if(fOTU_first_term == "dummy"){
    # For now just jump over those fOTUs with zero hits
    #fOTUsHMM <- bind_rows(fOTUsHMM,tibble())
    next
  }else{
    # If there were matches spread the -log_10 evalues for each
    # fOTU in one row
    fOTU <- fOTU %>%
      # Group all same HMMs together
      group_by(hmm_profile_id) %>%
      # Choose average of all e-values for each unique HMM
      summarise(avg_log_eval = mean(log_eval)) %>%
      add_column(fOTU_name = fOTU_first_term) %>%

      # Choose the largest value among the unique HMMs
      #summarise(max = max(log_eval)) %>%
      #add_column(fOTU_name = fOTU_first_term)

      # pivot_wider should work for tidyr v. 1.0.0
      # but I use 0.8.3, therefore I use spread() instead
      # pivot_wider(names_from = hmm_profile_id,
      #             values_from = log_eval)

      # Finally, spread the e-vals on one row
      spread(hmm_profile_id,avg_log_eval)
  }
  # Append the current fOTU data to one big tibble
  fOTUsHMM <- bind_rows(fOTUsHMM,fOTU)
}
```

Chapter 4

Exchange NA, Inf and -Inf with 0

```
is.na(fOTUsHMM) <- sapply(fOTUsHMM, is.infinite)
fOTUsHMM[is.na(fOTUsHMM)] <- 0
```

Chapter 5

Move fOTU names to row names and execute PCA

```
fOTUsHMM_pca <- fOTUsHMM %>%
  column_to_rownames(var = "fOTU_name") %>%
  prcomp(.,
         center = TRUE)
```

Chapter 6

Checkout the results of the PCA

```
summary(fOTUsHMM_pca)
```

```
## Importance of components:
##                             PC1      PC2      PC3      PC4      PC5
## Standard deviation     132.2534 68.03531 49.35914 47.70820 41.90192
## Proportion of Variance   0.2726  0.07213  0.03796  0.03547  0.02736
## Cumulative Proportion    0.2726  0.34467  0.38264  0.41810  0.44546
##                             PC6      PC7     PC8      PC9     PC10
## Standard deviation     39.03530 38.42849 37.5716 37.28472 36.21209
## Proportion of Variance  0.02374  0.02301  0.0220  0.02166  0.02043
## Cumulative Proportion   0.46920  0.49221  0.5142  0.53587  0.55631
##                            PC11     PC12     PC13     PC14    PC15
## Standard deviation     32.58043 31.62649 29.47653 29.08001 28.3280
## Proportion of Variance  0.01654  0.01559  0.01354  0.01318  0.0125
## Cumulative Proportion   0.57285  0.58843  0.60197  0.61515  0.6277
##                            PC16     PC17     PC18     PC19     PC20
## Standard deviation     28.11021 27.27425 26.27449 25.66268 24.84550
## Proportion of Variance  0.01231  0.01159  0.01076  0.01026  0.00962
## Cumulative Proportion   0.63996  0.65156  0.66231  0.67257  0.68219
##                            PC21    PC22     PC23     PC24     PC25
## Standard deviation     24.68215 24.1619 23.51678 23.13791 22.89763
## Proportion of Variance  0.00949  0.0091  0.00862  0.00834  0.00817
## Cumulative Proportion   0.69169  0.7008  0.70940  0.71774  0.72591
##                            PC26     PC27     PC28     PC29     PC30
## Standard deviation     22.54209 21.72888 21.47379 20.71456 20.31880
## Proportion of Variance  0.00792  0.00736  0.00719  0.00669  0.00643
```

```
## Cumulative Proportion   0.73383  0.74119  0.74837  0.75506  0.76149
##                            PC31     PC32     PC33     PC34     PC35
## Standard deviation     20.18789 19.97980 19.77073 18.87358 18.73493
## Proportion of Variance  0.00635  0.00622  0.00609  0.00555  0.00547
## Cumulative Proportion   0.76784  0.77406  0.78015  0.78570  0.79117
##                            PC36     PC37    PC38    PC39     PC40     PC41
## Standard deviation     18.35216 18.05607 17.9137 17.7398 17.27179 16.90352
## Proportion of Variance  0.00525  0.00508  0.0050  0.0049  0.00465  0.00445
## Cumulative Proportion   0.79642  0.80150  0.8065  0.8114  0.81605  0.82051
##                            PC42     PC43     PC44     PC45    PC46
## Standard deviation     16.65190 16.59387 16.27301 16.16877 16.0222
## Proportion of Variance  0.00432  0.00429  0.00413  0.00407  0.0040
## Cumulative Proportion   0.82483  0.82912  0.83324  0.83732  0.8413
##                            PC47     PC48     PC49     PC50     PC51
## Standard deviation     15.88432 15.55157 15.38475 15.17790 14.99940
## Proportion of Variance  0.00393  0.00377  0.00369  0.00359  0.00351
## Cumulative Proportion   0.84525  0.84902  0.85271  0.85629  0.85980
##                            PC52    PC53     PC54     PC55     PC56
## Standard deviation     14.59372 14.5597 14.35638 14.21161 14.07358
## Proportion of Variance  0.00332  0.0033  0.00321  0.00315  0.00309
## Cumulative Proportion   0.86312  0.8664  0.86963  0.87278  0.87587
##                            PC57     PC58     PC59     PC60     PC61
## Standard deviation     13.98450 13.79684 13.49591 13.44555 13.29861
## Proportion of Variance  0.00305  0.00297  0.00284  0.00282  0.00276
## Cumulative Proportion   0.87891  0.88188  0.88472  0.88754  0.89029
##                            PC62    PC63     PC64    PC65     PC66     PC67
## Standard deviation     13.11865 12.9275 12.88597 12.6677 12.53679 12.38289
## Proportion of Variance  0.00268  0.0026  0.00259  0.0025  0.00245  0.00239
## Cumulative Proportion   0.89297  0.8956  0.89816  0.9007  0.90311  0.90550
##                            PC68     PC69     PC70     PC71     PC72
## Standard deviation     12.36024 12.12626 11.90275 11.75965 11.49207
## Proportion of Variance  0.00238  0.00229  0.00221  0.00215  0.00206
## Cumulative Proportion   0.90788  0.91018  0.91238  0.91454  0.91660
##                           PC73     PC74     PC75     PC76     PC77
## Standard deviation     11.3275 11.25950 11.21988 11.08340 10.93554
## Proportion of Variance  0.0020  0.00198  0.00196  0.00191  0.00186
## Cumulative Proportion   0.9186  0.92057  0.92253  0.92445  0.92631
##                            PC78     PC79    PC80     PC81     PC82    PC83
## Standard deviation     10.88817 10.78407 10.4383 10.30526 10.27727 9.97812
## Proportion of Variance  0.00185  0.00181  0.0017  0.00165  0.00165 0.00155
## Cumulative Proportion   0.92816  0.92997  0.9317  0.93332  0.93497 0.93652
##                           PC84    PC85    PC86    PC87    PC88    PC89
## Standard deviation     9.85505 9.71454 9.63059 9.52854 9.43807 9.40537
## Proportion of Variance 0.00151 0.00147 0.00145 0.00141 0.00139 0.00138
## Cumulative Proportion  0.93803 0.93950 0.94095 0.94236 0.94375 0.94513
##                           PC90    PC91   PC92    PC93    PC94    PC95
```

```
## Standard deviation     9.34899 9.30698 9.1213 8.99382 8.96671 8.87664
## Proportion of Variance 0.00136 0.00135 0.0013 0.00126 0.00125 0.00123
## Cumulative Proportion  0.94649 0.94784 0.9491 0.95040 0.95165 0.95288
##                           PC96   PC97    PC98    PC99  PC100   PC101
## Standard deviation     8.84889 8.7921 8.69677 8.48411 8.4137 8.28001
## Proportion of Variance 0.00122 0.0012 0.00118 0.00112 0.0011 0.00107
## Cumulative Proportion  0.95410 0.9553 0.95648 0.95760 0.9587 0.95977
##                          PC102  PC103   PC104   PC105   PC106   PC107
## Standard deviation     8.13355 8.0115 7.90846 7.80160 7.71868 7.70733
## Proportion of Variance 0.00103 0.0010 0.00097 0.00095 0.00093 0.00093
## Cumulative Proportion  0.96081 0.9618 0.96278 0.96373 0.96466 0.96558
##                          PC108   PC109   PC110   PC111   PC112   PC113
## Standard deviation     7.57745 7.54595 7.50181 7.42419 7.35353 7.26900
## Proportion of Variance 0.00089 0.00089 0.00088 0.00086 0.00084 0.00082
## Cumulative Proportion  0.96648 0.96736 0.96824 0.96910 0.96994 0.97077
##                          PC114   PC115   PC116   PC117   PC118   PC119
## Standard deviation     7.1657 7.00915 6.94304 6.88961 6.81976 6.74789
## Proportion of Variance 0.0008 0.00077 0.00075 0.00074 0.00072 0.00071
## Cumulative Proportion  0.9716 0.97233 0.97308 0.97382 0.97455 0.97526
##                          PC120   PC121   PC122   PC123   PC124   PC125
## Standard deviation     6.63347 6.56315 6.50435 6.45755 6.32807 6.27955
## Proportion of Variance 0.00069 0.00067 0.00066 0.00065 0.00062 0.00061
## Cumulative Proportion  0.97594 0.97661 0.97727 0.97792 0.97855 0.97916
##                          PC126   PC127   PC128   PC129   PC130   PC131
## Standard deviation     6.1855 6.09281 6.04910 5.93077 5.78450 5.73490
## Proportion of Variance 0.0006 0.00058 0.00057 0.00055 0.00052 0.00051
## Cumulative Proportion  0.9798 0.98034 0.98091 0.98145 0.98198 0.98249
##                          PC132   PC133   PC134   PC135   PC136   PC137
## Standard deviation     5.72508 5.45415 5.39283 5.31520 5.26937 5.19411
## Proportion of Variance 0.00051 0.00046 0.00045 0.00044 0.00043 0.00042
## Cumulative Proportion  0.98300 0.98346 0.98392 0.98436 0.98479 0.98521
##                          PC138   PC139   PC140   PC141   PC142   PC143
## Standard deviation     5.16353 5.10980 5.02639 4.99259 4.93990 4.83252
## Proportion of Variance 0.00042 0.00041 0.00039 0.00039 0.00038 0.00036
## Cumulative Proportion  0.98562 0.98603 0.98642 0.98681 0.98719 0.98756
##                          PC144   PC145   PC146   PC147   PC148   PC149
## Standard deviation     4.80652 4.75977 4.62174 4.59725 4.54445 4.48196
## Proportion of Variance 0.00036 0.00035 0.00033 0.00033 0.00032 0.00031
## Cumulative Proportion  0.98792 0.98827 0.98860 0.98893 0.98925 0.98957
##                          PC150  PC151   PC152   PC153   PC154   PC155
## Standard deviation     4.43160 4.3543 4.26618 4.19748 4.15032 4.10529
## Proportion of Variance 0.00031 0.0003 0.00028 0.00027 0.00027 0.00026
## Cumulative Proportion  0.98987 0.9902 0.99045 0.99073 0.99099 0.99126
##                          PC156   PC157   PC158   PC159   PC160   PC161
## Standard deviation     4.10105 4.02575 3.97180 3.92081 3.89586 3.86421
## Proportion of Variance 0.00026 0.00025 0.00025 0.00024 0.00024 0.00023
```

```
## Cumulative Proportion  0.99152 0.99177 0.99202 0.99226 0.99249 0.99273
##                           PC162   PC163   PC164   PC165  PC166  PC167
## Standard deviation     3.80954 3.78049 3.67929 3.63640 3.6118 3.5875
## Proportion of Variance 0.00023 0.00022 0.00021 0.00021 0.0002 0.0002
## Cumulative Proportion  0.99295 0.99318 0.99339 0.99359 0.9938 0.9940
##                           PC168   PC169   PC170   PC171   PC172   PC173
## Standard deviation     3.49756 3.47330 3.44910 3.43466 3.34562 3.30931
## Proportion of Variance 0.00019 0.00019 0.00019 0.00018 0.00017 0.00017
## Cumulative Proportion  0.99419 0.99437 0.99456 0.99474 0.99492 0.99509
##                           PC174   PC175   PC176   PC177   PC178   PC179
## Standard deviation     3.27957 3.23234 3.19067 3.12284 3.11639 3.08484
## Proportion of Variance 0.00017 0.00016 0.00016 0.00015 0.00015 0.00015
## Cumulative Proportion  0.99526 0.99542 0.99558 0.99573 0.99588 0.99603
##                           PC180   PC181   PC182   PC183   PC184   PC185
## Standard deviation     3.04567 3.04000 2.95348 2.91897 2.91397 2.87321
## Proportion of Variance 0.00014 0.00014 0.00014 0.00013 0.00013 0.00013
## Cumulative Proportion  0.99617 0.99632 0.99645 0.99659 0.99672 0.99685
##                           PC186   PC187   PC188   PC189   PC190  PC191
## Standard deviation     2.80510 2.76727 2.67537 2.63741 2.62961 2.5764
## Proportion of Variance 0.00012 0.00012 0.00011 0.00011 0.00011 0.0001
## Cumulative Proportion  0.99697 0.99709 0.99720 0.99731 0.99742 0.9975
##                           PC192  PC193  PC194   PC195   PC196   PC197
## Standard deviation     2.5604 2.5099 2.4878 2.45680 2.41467 2.37648
## Proportion of Variance 0.0001 0.0001 0.0001 0.00009 0.00009 0.00009
## Cumulative Proportion  0.9976 0.9977 0.9978 0.99791 0.99800 0.99809
##                           PC198   PC199   PC200   PC201   PC202   PC203
## Standard deviation     2.31245 2.28498 2.25082 2.22246 2.19272 2.15384
## Proportion of Variance 0.00008 0.00008 0.00008 0.00008 0.00007 0.00007
## Cumulative Proportion  0.99817 0.99826 0.99833 0.99841 0.99849 0.99856
##                           PC204   PC205   PC206   PC207   PC208   PC209
## Standard deviation     2.10798 2.09091 2.07066 2.04050 2.03510 1.96520
## Proportion of Variance 0.00007 0.00007 0.00007 0.00006 0.00006 0.00006
## Cumulative Proportion  0.99863 0.99870 0.99876 0.99883 0.99889 0.99895
##                           PC210   PC211   PC212   PC213   PC214   PC215
## Standard deviation     1.93124 1.88442 1.85902 1.83575 1.81009 1.77746
## Proportion of Variance 0.00006 0.00006 0.00005 0.00005 0.00005 0.00005
## Cumulative Proportion  0.99901 0.99907 0.99912 0.99917 0.99922 0.99927
##                           PC216   PC217   PC218   PC219   PC220   PC221
## Standard deviation     1.71386 1.68086 1.66189 1.65646 1.61684 1.54967
## Proportion of Variance 0.00005 0.00004 0.00004 0.00004 0.00004 0.00004
## Cumulative Proportion  0.99932 0.99936 0.99941 0.99945 0.99949 0.99953
##                           PC222   PC223   PC224   PC225   PC226   PC227
## Standard deviation     1.52385 1.50528 1.48662 1.46114 1.39884 1.35834
## Proportion of Variance 0.00004 0.00004 0.00003 0.00003 0.00003 0.00003
## Cumulative Proportion  0.99956 0.99960 0.99963 0.99967 0.99970 0.99972
##                           PC228   PC229   PC230   PC231   PC232   PC233
```

```
## Standard deviation     1.30189 1.26736 1.26198 1.23665 1.17053 1.12286
## Proportion of Variance 0.00003 0.00003 0.00002 0.00002 0.00002 0.00002
## Cumulative Proportion  0.99975 0.99978 0.99980 0.99982 0.99985 0.99987
##                           PC234   PC235   PC236   PC237   PC238   PC239
## Standard deviation     1.04115 1.02192 1.00972 0.95880 0.92698 0.86118
## Proportion of Variance 0.00002 0.00002 0.00002 0.00001 0.00001 0.00001
## Cumulative Proportion  0.99988 0.99990 0.99991 0.99993 0.99994 0.99995
##                           PC240   PC241   PC242  PC243  PC244  PC245  PC246
## Standard deviation     0.73246 0.68825 0.65293 0.5296 0.5187 0.4474 0.4083
## Proportion of Variance 0.00001 0.00001 0.00001 0.0000 0.0000 0.0000 0.0000
## Cumulative Proportion  0.99996 0.99997 0.99998 1.0000 1.0000 1.0000 1.0000
##                          PC247  PC248  PC249  PC250  PC251 PC252     PC253
## Standard deviation     0.3931 0.3593 0.3142 0.3094 0.2912 0.216 5.873e-14
## Proportion of Variance 0.0000 0.0000 0.0000 0.0000 0.0000 0.000 0.000e+00
## Cumulative Proportion  1.0000 1.0000 1.0000 1.0000 1.0000 1.000 1.000e+00
```

Doesn't seem like PCA will be very useful.

Chapter 7

Visualise the results

The following graph depicts how large percentage of bins in fOTU have at least one hit from viral HMM profiles with e-value less than 0.01.

```
qplot(hit_percentage, data = hits, binwidth = 0.01)
```

Chapter 8

Not used at the moment

This following is not run because they are not of interest at the moment.

```
fOTUbinNums <- read_csv(file = "../analyses/numBinsfOTU.csv",
                          col_types = "ci",
                          col_names = c("fOTU_name",
                                        "num_bins"))
egg_nog_cats <-
  read_tsv(file = "../data/annotations/10239_annotations.tsv",
           col_types = "fcfc",
           col_names = c("taxid","hmm_profile_id",
                         "egg_nog_category",
                         "hmm_description")) %>%
  # Drop taxid because it's uninteresting
  select(.,-taxid)

num_seqs <- read_csv(file = "../analyses/numSeqsBin.csv",
                          col_types = "ci")
```

```
# Make a big tibble out of these three tibbles by left joining
fOTU <- left_join(fOTU,
                  egg_nog_cats,
                  by = "hmm_profile_id") %>%
  left_join(.,
            fOTUbinNums,
            by = "fOTU_name") %>%
  # Remove values that were outside inclusion threshold
  filter(.,inside_inclusion_threshold) %>%
  # Drop inside_inclusion_threshold now that it has done its duty
  select(.,-inside_inclusion_threshold)
```

```
  # Find how many hits there are to each bin
  numHits <- fOTU %>%
    count(.,Target_Bin_id) %>%
    rename(.,Num_hits = n)

  # Add num hits beside each bin
  fOTU <- left_join(fOTU, numHits, by = "Target_Bin_id")

  # Add total number of sequences in each bin
  fOTU <- left_join(fOTU, num_seqs, by = "Target_Bin_id")

  # Count how many unique hits to bins there are
  fOTU_vec_len <- fOTU %>%
    distinct(.,Target_Bin_id) %>%
    pull(Target_Bin_id) %>%
    length()

  # Grab the total number of bins in the fOTU
  fOTU_tot_num_bins <- fOTU %>%
    pull(num_bins)
  fOTU_tot_num_bins <- fOTU_vec_len/fOTU_tot_num_bins[1]

  # Append the value to a vector
#  hit_percentage <- c(hit_percentage,fOTU_tot_num_bins)

#}

# Create finally a tibble from the vector
#hits <- tibble(hit_percentage)
```

Chapter 9

Session info

```
sessionInfo()
```

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Debian GNU/Linux 9 (stretch)
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib/libopenblasp-r0.2.19.so
##
## locale:
##  [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8
##  [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C
##  [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C
## [10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] forcats_0.4.0   stringr_1.4.0   dplyr_0.8.3     purrr_0.3.2
## [5] readr_1.3.1     tibble_2.1.3    ggplot2_3.2.1   tidyverse_1.2.1
## [9] tidyr_0.8.3
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.2       cellranger_1.1.0 pillar_1.4.2     compiler_3.6.1
##  [5] tools_3.6.1      zeallot_0.1.0    digest_0.6.20    lubridate_1.7.4
##  [9] jsonlite_1.6     evaluate_0.14    nlme_3.1-140     gtable_0.3.0
## [13] lattice_0.20-38  pkgconfig_2.0.2  rlang_0.4.0      cli_1.1.0
## [17] rstudioapi_0.10  yaml_2.2.0       haven_2.1.1      xfun_0.9
```

```
## [21] withr_2.1.2      xml2_1.2.2       httr_1.4.1       knitr_1.24
## [25] vctrs_0.2.0      hms_0.5.1        generics_0.0.2   grid_3.6.1
## [29] tidyselect_0.2.5 glue_1.3.1       R6_2.4.0         readxl_1.3.1
## [33] rmarkdown_1.15   bookdown_0.13    modelr_0.1.5     magrittr_1.5
## [37] backports_1.1.4  scales_1.0.0     htmltools_0.3.6  rvest_0.3.4
## [41] assertthat_0.2.1 colorspace_1.4-1 stringi_1.4.3    lazyeval_0.2.2
## [45] munsell_0.5.0    broom_0.5.2      crayon_1.3.4
```

Chapter 10

References