Longer ORF PDB search and results

Mirte Ciz Marieke Kuijpers

02/03/2022

The code in this document is made to be useful with either the long or the short ORF, but in the set-up below the sequence to use is set to the long ORF.

```
# Set-up
library("bio3d")
library("ggplot2")
library("ggrepel")
library("msa")
library("bio3d.view")

# Load sequence of POI
seqL <- read.fasta("long.ORF.fa")
seqS <- read.fasta("short.ORF.fa")

#Choose which sequence to use
seq <- seqL
```

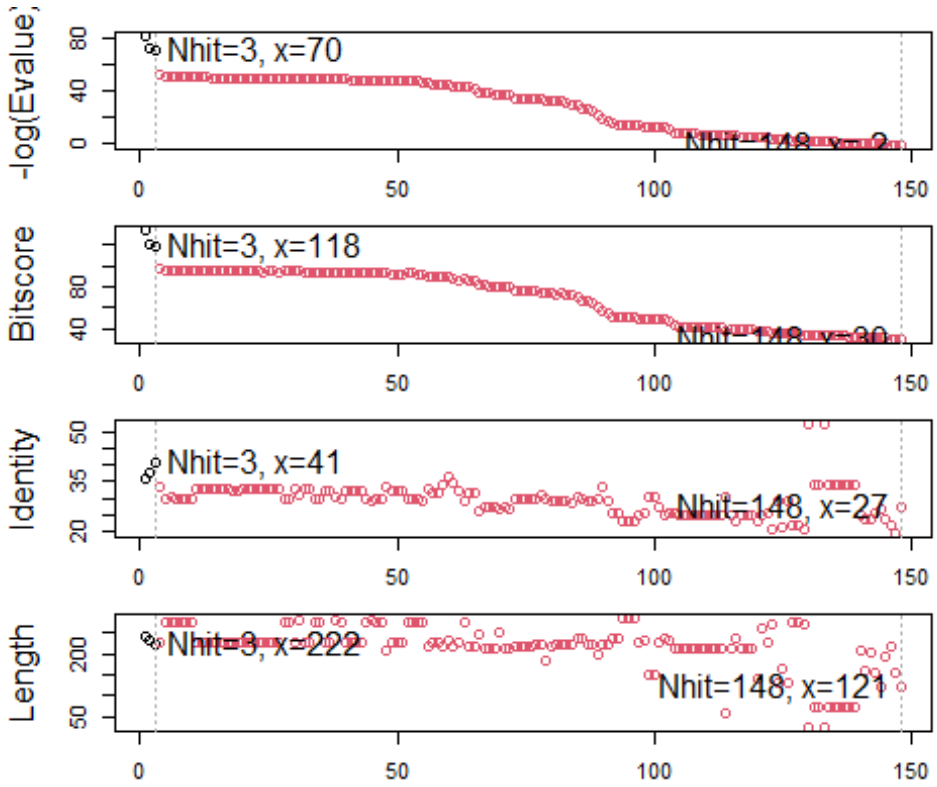
After set up the blast search can be completed and the summary statistics of this search can be plotted.

```
# Blast search
blast <- blast.pdb(seq, database = "pdb")
```

```
##  Searching ... please wait (updates every 5 seconds) RID = 1ZJ94FWD01R
##
.....................................................................
.....................................................................
.....................................................................
.....................................................
##  Reporting 148 hits
```

```
# Plot summary statistics of results
hits <- plot(blast)
```

```
##   * Possible cutoff values:    70 -2
##              Yielding Nhits:    3 148
##
##   * Chosen cutoff value of:    70
##              Yielding Nhits:    3
```




```
# Print the IDs of the hits above the threshold
hit.IDs <- hits$pdb.id
hit.IDs

## [1] "7BMH_A" "5AWZ_A" "6GYH_A"
```

There are 3 hits that pass the statistical threshold, namely: 7BMH_A, 5AWZ_A, 6GYH_A. More information can be found on these by interrogating the blast results.

```
# Show the hit table for the top hits which pass the threshold
head(blast$hit.tbl, n = length(hit.IDs))

##        queryid subjectids identity alignmentlength mismatches gapopens
q.start
## 1 Query_540941     7BMH_A   35.685             241        149         3
42
## 2 Query_540941     5AWZ_A   37.662             231        115         5
39
## 3 Query_540941     6GYH_A   40.541             222        122         4
46
##   q.end s.start s.end   evalue bitscore positives mlog.evalue pdb.id
acc
## 1   276      51   291 6.05e-36      133     51.45    81.09301 7BMH_A
7BMH_A
## 2   258      16   228 6.87e-32      120     49.78    71.75556 5AWZ_A
5AWZ_A
```

```
## 3   266      14   226 2.73e-31      118     55.86    70.37584 6GYH_A
6GYH_A
```

We can also download these PDB files, annotate them for more information and align them with our sequence to get an overview of sequence alignment.

```
# Download related PDB files
files <- get.pdb(hits$pdb.id, path="pdbs", split=TRUE, gzip=TRUE)
```

```
##   |
|                                                                      |   0%
|
|=======================                                               |  33%
|
|===============================================                       |  67%
|
|======================================================================| 100%
```

```
# Align related PDBs
pdbs <- pdbaln(files, fit = TRUE, exefile="msa")
```

```
## Reading PDB files:
## pdbs/split_chain/7BMH_A.pdb
## pdbs/split_chain/5AWZ_A.pdb
## pdbs/split_chain/6GYH_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
## .   PDB has ALT records, taking A only, rm.alt=TRUE
## .   PDB has ALT records, taking A only, rm.alt=TRUE
## .
##
## Extracting sequences
##
## pdb/seq: 1   name: pdbs/split_chain/7BMH_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 2   name: pdbs/split_chain/5AWZ_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 3   name: pdbs/split_chain/6GYH_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
```

```
# Vector containing PDB codes for figure axis
ids <- basename.pdb(pdbs$id)

# Annotate hits for more information on the hits
anno <- pdb.annotate(ids)

# Find the organisms these PDB hits come from
unique(anno$source)
```

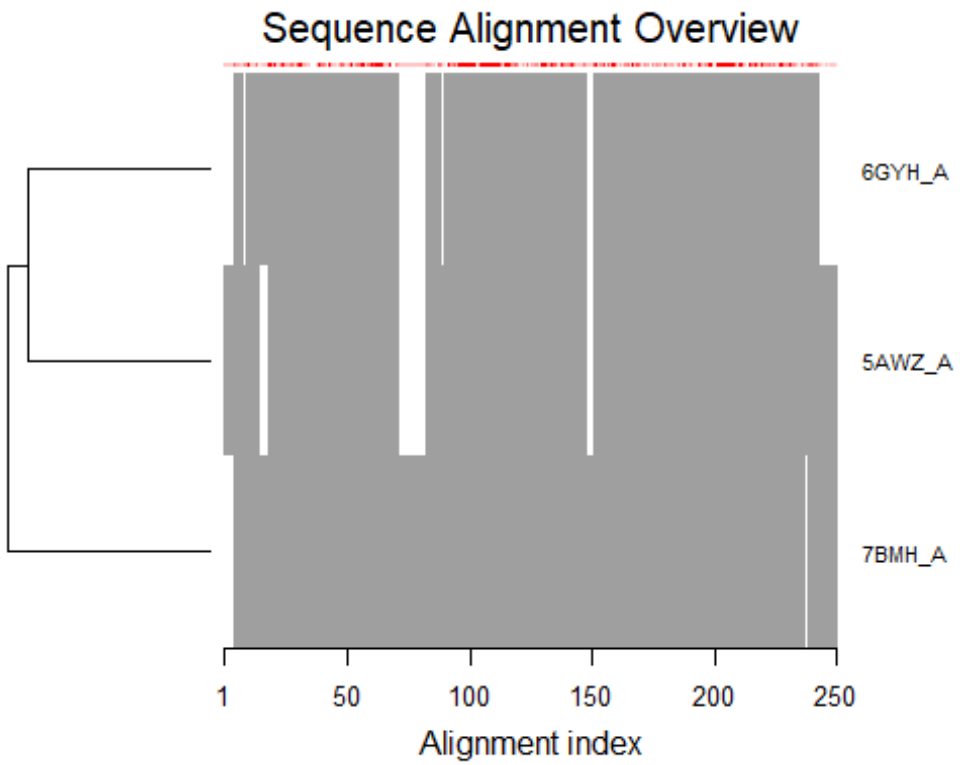
```
## [1] "Leptosphaeria maculans"         "Acetabularia acetabulum"
## [3] "Coccomyxa subellipsoidea C-169"
```

```
# View more information on the hits
anno

##        structureId chainId macromoleculeType chainLength
experimentalTechnique
## 7BMH_A        7BMH       A           Protein         324
X-ray
## 5AWZ_A        5AWZ       A           Protein         244
X-ray
## 6GYH_A        6GYH       A           Protein         236
X-ray
##        resolution scopDomain
pfam
## 7BMH_A       2.20       <NA> Bacteriorhodopsin-like protein
(Bac_rhodopsin)
## 5AWZ_A       1.57       <NA> Bacteriorhodopsin-like protein
(Bac_rhodopsin)
## 6GYH_A       2.00       <NA> Bacteriorhodopsin-like protein
(Bac_rhodopsin)
##                                ligandId
## 7BMH_A                    LFA (22),OLA (3)
## 5AWZ_A RET,OLB,D12 (2),D10 (3),OCT (2),C14
## 6GYH_A                     RET,CLR,OLB (4)
##
ligandName
## 7BMH_A
EICOSANE (22),OLEIC ACID (3)
## 5AWZ_A RETINAL,(2S)-2,3-dihydroxypropyl (9Z)-octadec-9-enoate,DODECANE
(2),DECANE (3),N-OCTANE (2),TETRADECANE
## 6GYH_A                                    RETINAL,CHOLESTEROL,(2S)-2,3-
dihydroxypropyl (9Z)-octadec-9-enoate (4)
##                                 source
## 7BMH_A         Leptosphaeria maculans
## 5AWZ_A        Acetabularia acetabulum
## 6GYH_A Coccomyxa subellipsoidea C-169
##
structureTitle
## 7BMH_A                        Crystal structure of a light-driven proton
pump LR (Mac) from Leptosphaeria maculans
## 5AWZ_A Crystal Structure of the Cell-Free Synthesized Membrane Protein,
Acetabularia Rhodopsin I, at 1.57 angstrom
## 6GYH_A                     Crystal structure of the light-driven proton
pump Coccomyxa subellipsoidea Rhodopsin CsR
##                                                                  citation
rObserved
## 7BMH_A                      Zabelskii, D., et al. Commun Biol (2021)
0.23840
## 5AWZ_A Furuse, M., et al. Acta Crystallogr D Biol Crystallogr (2015)
0.17760
## 6GYH_A                            Fudim, R., et al. Sci Signal (2019)
```

```
0.19398
##          rFree   rWork spaceGroup
## 7BMH_A 0.28470 0.23610 P 21 21 21
## 5AWZ_A 0.19410 0.17690    C 1 2 1
## 6GYH_A 0.22493 0.19231        H 3

# Draw schematic alignment
plot(pdbs, labels=ids)
```

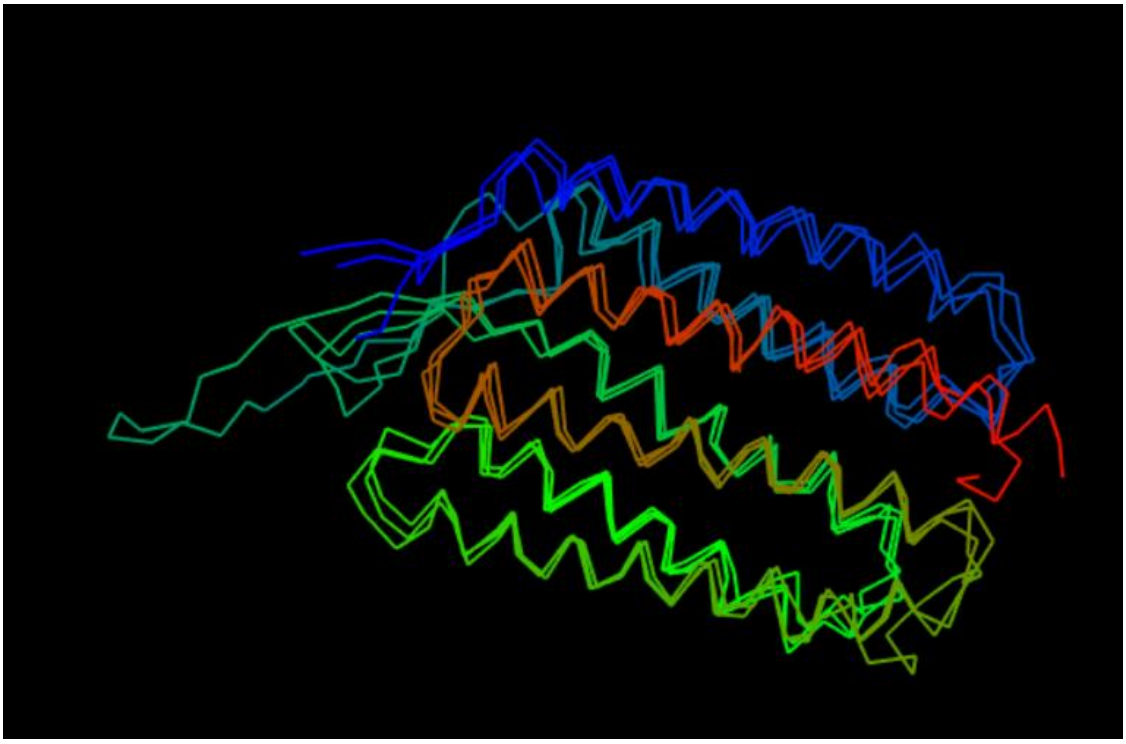


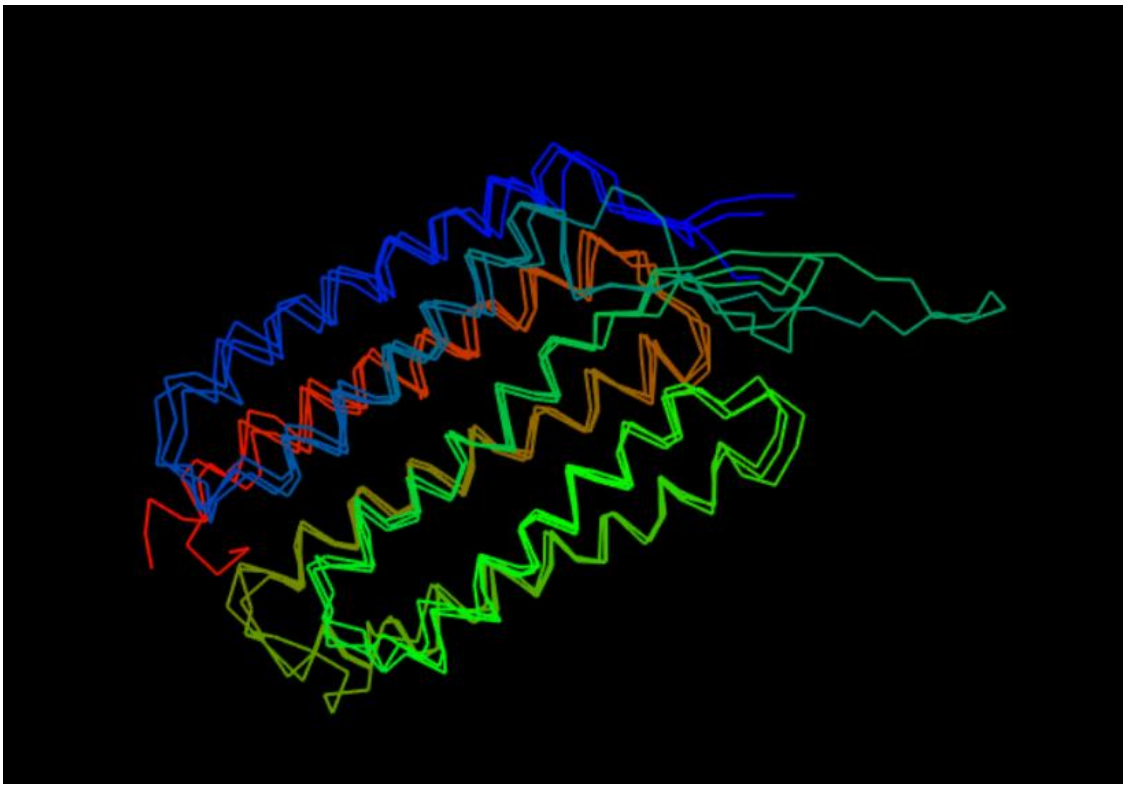
We can also plot the three structures we have found as follows:

```
# Set up
library(bio3d.view)
library(rgl)

# Plot
#view.pdbs(pdbs)
```

The `View.pdbs()` function brings up an interactive viewer, which cannot be directly viewed in the markdown document, so instead two screen-shots of this have been inserted.





With more proteins it could be interesting to plot variability, or even do PCA using the amino acid position data, but with only three proteins this is not useful.

Session Information

```
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22000)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.1252
## [2] LC_CTYPE=English_United Kingdom.1252
## [3] LC_MONETARY=English_United Kingdom.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.1252
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] rgl_0.108.3           bio3d.view_0.1.0.9000 msa_1.26.0
##  [4] Biostrings_2.62.0     GenomeInfoDb_1.30.1   XVector_0.34.0
##  [7] IRanges_2.28.0        S4Vectors_0.32.3      BiocGenerics_0.40.0
## [10] ggrepel_0.9.1         ggplot2_3.3.5         bio3d_2.4-3.9000
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.2       xfun_0.29              purrr_0.3.4
##  [4] colorspace_2.0-2       vctrs_0.3.8            generics_0.1.2
##  [7] htmltools_0.5.2        yaml_2.2.2             utf8_1.2.2
## [10] rlang_1.0.1            pillar_1.7.0           glue_1.6.1
## [13] withr_2.4.3            GenomeInfoDbData_1.2.7 lifecycle_1.0.1
## [16] stringr_1.4.0          zlibbioc_1.40.0        munsell_0.5.0
## [19] gtable_0.3.0           htmlwidgets_1.5.4      evaluate_0.15
## [22] knitr_1.37             extrafont_0.17         fastmap_1.1.0
## [25] curl_4.3.2             parallel_4.1.2         fansi_1.0.2
## [28] Rttf2pt1_1.3.10        highr_0.9              Rcpp_1.0.8
## [31] scales_1.1.1           jsonlite_1.8.0         digest_0.6.29
## [34] stringi_1.7.6          dplyr_1.0.8            grid_4.1.2
## [37] cli_3.2.0              tools_4.1.2            bitops_1.0-7
## [40] magrittr_2.0.2         RCurl_1.98-1.6         tibble_3.1.6
## [43] extrafontdb_1.0        crayon_1.5.0           pkgconfig_2.0.3
## [46] ellipsis_0.3.2         httr_1.4.2             rmarkdown_2.11
## [49] rstudioapi_0.13        R6_2.5.1               compiler_4.1.2
```