The journey of Multiple Sequence Alignment

Diving in the Notum

Constantinos Yeles (Konstantinos Geles)

First version: Tue 26/07/2021, Latest update: Mon 26/Jul/2021

Contents

Blast alignment

First we need to download the fasta sequence and the database from dropbox? Maybe can be uploaded later to zenodo (probably I will have more context to add later).

We will use local blast run on docker as the link wasn't working.

```
# get the docker and run
docker pull ncbi/blast
docker run --rm -ti -v $(pwd):/home/my_data  ncbi/blast
# move inside /home dir of the container
cd /home

# create a folder for the ouput of the blast database and unzip the fasta sequences
mkdir my_data/some_species_blast_db
gunzip my_data/oma_db.gz

# make a blast database
makeblastdb  -dbtype "prot" -in "my_data/oma_db" -input_type "fasta" -out "my_data/some_species_blast_d
```

```
# blast the sequence of Notum
blastp -query "my_data/dm_notum.fasta"  -db "my_data/some_species_blast_db/blast_oma.db" -out "my_data/
```

Import results into R

print an example of the results

```
notum_blst %>% arrange(evalue)
```

```
## GRanges object with 197 ranges and 26 metadata columns:
##                       seqnames    ranges strand |    query_id query_length
##                          <Rle> <IRanges>  <Rle> | <character>    <numeric>
##     [1]              Dm|Notum     1-671      + |    Dm|Notum          671
##     [2]      Bf|XP_002599183.1    86-411      + |    Dm|Notum          671
##     [3]           La|919081898    67-400      + |    Dm|Notum          671
##     [4]           Sk|269784925    68-450      + |    Dm|Notum          671
##     [5]          Ct|ELT91091.1    48-379      + |    Dm|Notum          671
##     ...                    ...       ...    ... .         ...          ...
##   [193]            sb|3460350|   526-652      + |    Dm|Notum          671
##   [194]      Ta|XP_002109594.1    32-146      + |    Dm|Notum          671
##   [195]           Cg|762136134 9401-9532      + |    Dm|Notum          671
##   [196]          d_gig|g110.t1   408-437      + |    Dm|Notum          671
##   [197] a_aur|scaffold226.g1.. 7906-8083      + |    Dm|Notum          671
##                     subject_id subject_length   q_start     q_end   s_start
##                    <character>      <numeric> <numeric> <numeric> <numeric>
##     [1]              Dm|Notum            671         1       671         1
##     [2]      Bf|XP_002599183.1            487        88       412        86
##     [3]           La|919081898            559        88       418        67
##     [4]           Sk|269784925            508        88       470        68
##     [5]          Ct|ELT91091.1            462        88       418        48
##     ...                    ...            ...       ...       ...       ...
##   [193]            sb|3460350|           3301       435       564       526
##   [194]      Ta|XP_002109594.1           1016       412       538        32
##   [195]           Cg|762136134           9959       443       579      9401
##   [196]          d_gig|g110.t1            467       598       627       408
##   [197] a_aur|scaffold226.g1..           8310       417       593      7906
##             s_end             query_seq            subject_seq    evalue
##         <numeric>           <character>            <character> <numeric>
##     [1]       671 MAVEQIDKMAAKAGEATNKW.. MAVEQIDKMAAKAGEATNKW..  0.00e+00
##     [2]       411 LKRANLANTSITCNDGSHAG.. MKLHKLRNTSVTCNDGSPAG.. 4.55e-121
##     [3]       400 LKRANLANTSITCNDGSHAG.. LKRHFLTNRTVTCNDGSPAG.. 1.01e-119
##     [4]       450 LKRANLANTSITCNDGSHAG.. MKLRYLENTTVTCNDGSPAG.. 2.01e-113
##     [5]       379 LKRANLANTSITCNDGSHAG.. MKRHFIRNPSVTCNDGSKAG.. 7.05e-108
##     ...       ...                   ...                    ...       ...
##   [193]       652 EHANNQRHQRHRQRLQRQKH.. EQRTENREQRTENREQRTEN..       5.2
##   [194]       146 STRSRRHDKLKRSTEPSTAV.. SERRRNHDDDKGSSRHARKE..       6.5
##   [195]      9532 QRHRQRLQRQKHNNVAQSGG.. ERKRKMLEKYKQLEEELEAE..       6.9
##   [196]       437 CGLRLLERCSWPQCNHSCPT.. CKSALIDDCSQPLCNSLCPA..       7.2
##   [197]      8083 RHDKLKRSTEPSTAVSHPEH.. RKRREKKKGEEVEHGSDKEE..       8.4
```

```
##         bit_score     score alignment_length percent_identity identical
##         <numeric> <numeric>        <numeric>        <numeric> <numeric>
##     [1]      1400      3625              671          100.000       671
##     [2]       373       957              333           53.453       178
##     [3]       372       954              339           51.327       174
##     [4]       353       907              396           44.697       177
##     [5]       338       866              338           48.817       165
##     ...       ...       ...              ...              ...       ...
##   [193]      35.0        79              130           23.846        31
##   [194]      34.7        78              127           29.921        38
##   [195]      34.7        78              141           27.660        39
##   [196]      34.3        77               30           46.667        14
##   [197]      34.3        77              182           19.231        35
##         mismatches positives percent_positives query_sbjct_frames query_frame
##          <numeric> <numeric>         <numeric>        <character>   <numeric>
##     [1]          0       671            100.00                1/1           1
##     [2]        140       231             69.37                1/1           1
##     [3]        152       238             70.21                1/1           1
##     [4]        193       244             61.62                1/1           1
##     [5]        160       224             66.27                1/1           1
##     ...        ...       ...               ...                ...         ...
##   [193]         96        57             43.85                1/1           1
##   [194]         77        62             48.82                1/1           1
##   [195]         89        70             49.65                1/1           1
##   [196]         16        17             56.67                1/1           1
##   [197]        138        74             40.66                1/1           1
##         sbjct_frame subject_strand percent_query_coverage_per_subject
##           <numeric>    <character>                          <numeric>
##     [1]           1            N/A                                100
##     [2]           1            N/A                                 59
##     [3]           1            N/A                                 60
##     [4]           1            N/A                                 66
##     [5]           1            N/A                                 63
##     ...         ...            ...                                ...
##   [193]           1            N/A                                 23
##   [194]           1            N/A                                 19
##   [195]           1            N/A                                 20
##   [196]           1            N/A                                 62
##   [197]           1            N/A                                 29
##         percent_query_coverage_per_hsp percent_query_coverage_per_uniq_subject
##                              <numeric>                               <character>
##     [1]                            100                                       N/A
##     [2]                             48                                       N/A
##     [3]                             49                                       N/A
##     [4]                             57                                       N/A
##     [5]                             49                                       N/A
##     ...                            ...                                       ...
##   [193]                             19                                       N/A
##   [194]                             19                                       N/A
##   [195]                             20                                       N/A
##   [196]                              4                                       N/A
##   [197]                             26                                       N/A
##   -------
##   seqinfo: 110 sequences from an unspecified genome; no seqlengths
```

get the 10 top hits identifiers

```
top10_seqs <- notum_blst %>% arrange(evalue) %>% head(11) %>%  .$subject_id
```

import fasta db

import the sequences and filter them for the 10 best hits from the previous results in order to make the multiple sequence alignments

```
suppressPackageStartupMessages({
library(Biostrings)
  })
fasta_db <- readAAStringSet("oma_db")

# I didn't consider that the names will be cut in the db creation thus I have to manipulate them to wor
names_fst_db <- fasta_db  %>% names %>% str_remove(" .+")

# filter for the top 10
fasta_db_fil <- fasta_db[names_fst_db %in% top10_seqs]
```

Multiple Sequence Alignment

I found two packages in R implementing various algorithms. One is msa and the other DECIPHER for now I will try msa

```
suppressPackageStartupMessages({
library(msa)
  })
myFirstAlignment <- msa(fasta_db_fil)
```

```
## use default substitution matrix
```

print the MSA

```
myFirstAlignment
```

```
## CLUSTAL 2.1
##
## Call:
##    msa(fasta_db_fil)
##
## MsaAAMultipleAlignment with 11 rows and 744 columns
##      aln                                                  names
##  [1] ----------------------ML...DDLVRMLTS---------------- Bf|XP_002599183.1...
##  [2] ----------------MGRGVRVLL...SELLGMLSNGS-------------- Hs|NOTUM gi|76799...
```

```
##  [3] --------------------MLLFL...QTMAMMMEG---------------- Sk|269784925 ref|...
##  [4] -------------------------...PTLQAMDHDVLLRLLVKNSR----- Ct|ELT91091.1 ELT...
##  [5] -------------------------...KTLQTMDHAKLIKLLMEQE------ Pa|g11818.t1
##  [6] -------------------MWLSFI...ATMQAMDHETLLKILTQQQ------ La|919081898 ref|...
##  [7] ---------------------MTKI...STVNAMNANTISTLLLDTFRQKRPQ Cg|762161189 ref|...
##  [8] MAVEQIDKMAAKAGEATNKWIKPQQ...HTLNNMERTELVNMLTQQAN----- Dm|Notum NP_73009...
##  [9] ----------------------MK...MMLATIEPHLILQMLLST------- Xb|g3677.t1
## [10] ----------------------MLQ...QVDRGFQ------------------- Ep|XP_020912206.1...
## [11] --------------------MDTLW...NLNRD--------------------- Nv|NVE7485
##  Con --------------------????...?TL??M????????L???------- Consensus
```