Homework 2

Björn Bebensee (2019–21343)
Machine Learning in Bioinformatics

June 10, 2020

1 Multiple Sequence Alignment using MUSCLE

First, in order to do further analysis with the given ten SARS-CoV-2 genome sequences and to generate the pHMM, we have to obtain a multiple sequence alignment. In order to align these sequences we can use MUSCLE [1]. Using MUSCLE we simply input the source sequences in .fasta format and run the command seen in Figure 1.

```
haean 福 ~/Documents/uni/Lecture Notes/2020 Spring/Machine Learning in Bioinformatics/HW2 ➤ bf68553|master⚡
[9044 ± : muscle -in mlbio_input_seq.fasta -out input_seq.msa -clwstrict

MUSCLE v3.8.1551 by Robert C. Edgar

http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

mlbio_input_seq 10 seqs, lengths min 29724, max 29903, avg 29813
00:00:00      3 MB(0%)  Iter   1  100.00%  K-mer dist pass 1
00:00:00      3 MB(0%)  Iter   1  100.00%  K-mer dist pass 2
00:02:58   1111 MB(6%)  Iter   1  100.00%  Align node
00:02:58   1111 MB(6%)  Iter   1  100.00%  Root alignment
00:04:56   1120 MB(7%)  Iter   2  100.00%  Refine tree
00:04:56   1120 MB(7%)  Iter   2  100.00%  Root alignment
00:04:56   1120 MB(7%)  Iter   2  100.00%  Root alignment
00:09:56   1120 MB(7%)  Iter   3  100.00%  Refine biparts
```

Figure 1: Command used to align the sequences

The output of MUSCLE is a multiple sequence alignment that can be seen in Figure 2, here in CLUSTAL W format. We can see that these sequences have now been aligned. Interestingly we can notice right away that the five sequences from Germany, namely MT358639, MT358640, MT358641, MT358642 and MT358643, have been perfectly aligned to each other at the beginning while the beginning of the five sequences from the USA, namely MT370833, MT370834, MT370835, MT370836 and MT370837, have been perfectly aligned to each other as well but offset by a little from the beginning of the sequences from Germany, meaning the American sequences do not contain this part at the beginning.

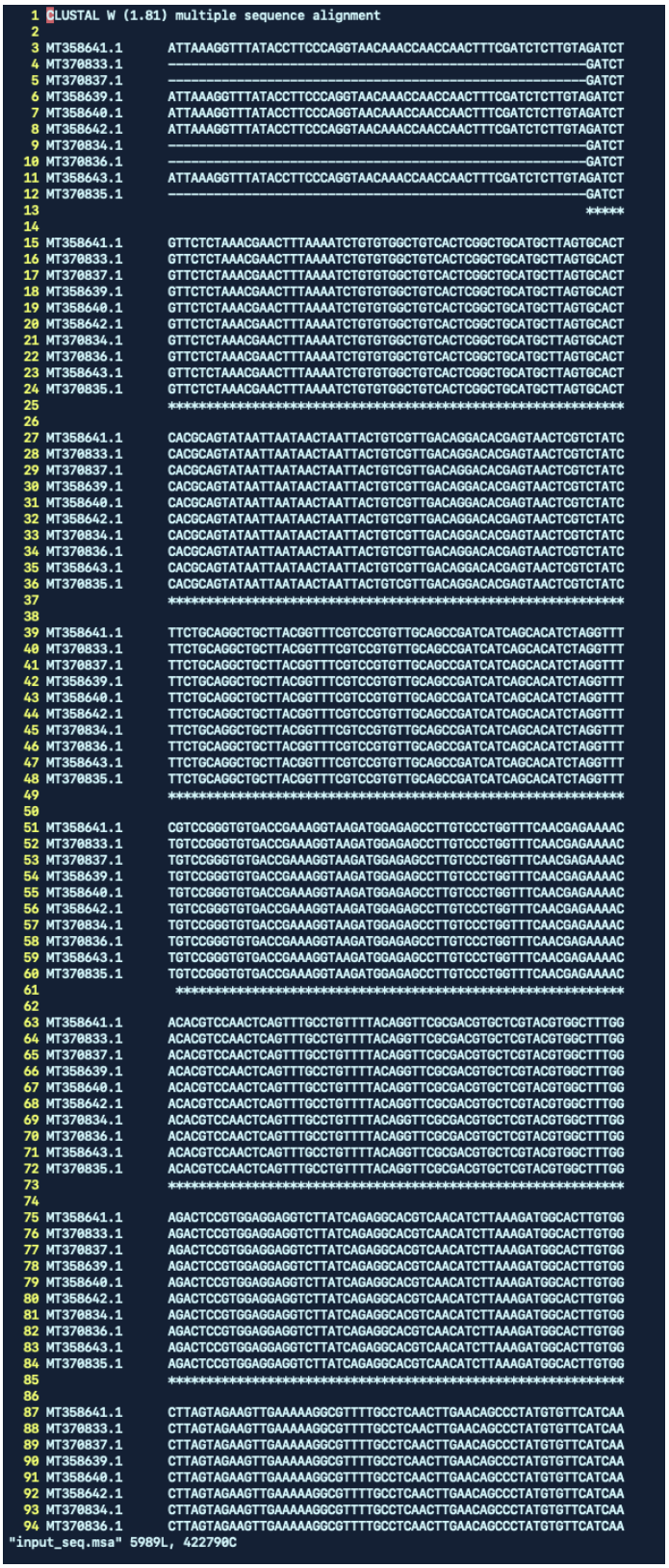


Figure 2: Resulting multiple sequence alignment

2 Building a profile HMM using HMMER

Now that we have obtained a multiple sequence alignment for our sequences, we can run the `hmmbuild` command in `HMMER` [2] as seen in Figure 3. As we can see there are ten aligned sequences in the input multiple sequence alignment (`nseq`) with 29 903 aligned columns (`alen`). The resulting output profile by HMM has the same amout of consensus positions (`mlen`) and a relative entropy per position of 0.622 bits (`re/pos`).

```
haean 福 ~/Documents/uni/Lecture Notes/2020 Spring/Machine Learning in Bioinformatics/HW2 ➤ bf68553|master⚡
[9058 ± : hmmbuild covid.hmm input_seq.msa
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.3 (Nov 2019); http://hmmer.org/
# Copyright (C) 2019 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# input alignment file:             input_seq.msa
# output HMM file:                  covid.hmm
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

# idx name                 nseq  alen  mlen     W eff_nseq re/pos description
#---- -------------------- ----- ----- ----- ----- -------- ------ -----------
1     input_seq               10 29903 29903 31848     1.70  0.622

# CPU time: 83.77u 0.20s 00:01:23.97 Elapsed: 00:01:24.34
```

Figure 3: Command to build the HMM

The resulting profile hidden Markov model can be seen in Figure 4. In this file we can see the metadata section and below it the actual HMM parameters. More specifically, these model parameters are negative log probabilities that describe the transition probabilities between the different states (that can be seen at the top of the parameter section). Each match state (enumerated on the left side from 1 to `MAXL`, here 29 903) contains three lines: match emission probabilities, insert emission probabilities and finally transition probabilities between the states.

```
HMMER3/f [3.3 | Nov 2019]
NAME  input_seq
LENG  29903
MAXL  31848
ALPH  DNA
RF    no
MM    no
CONS  yes
CS    no
MAP   yes
DATE  Mon Jun  8 00:16:03 2020
NSEQ  10
EFFN  1.704102
CKSUM 3798645073
STATS LOCAL MSV      -18.4306  0.69322
STATS LOCAL VITERBI  -18.9873  0.69322
STATS LOCAL FORWARD  -11.0465  0.69322
HMM          A        C        G        T
            m->m     m->i     m->d     i->m     i->i     d->m     d->d
  COMPO   1.26776  1.56207  1.47607  1.27187
          1.38629  1.38629  1.38629  1.38629
          0.31396  3.66461  1.41124  1.46634  0.26236  0.00000        *
      1   0.62298  2.01106  1.78035  1.82493      1 a - - -
          1.38629  1.38629  1.38629  1.38629
          0.06777  3.41842  3.41842  1.46634  0.26236  1.54840  0.23900
      2   1.85387  1.73195  1.88786  0.66352      2 t - - -
          1.38629  1.38629  1.38629  1.38629
          0.06777  3.41842  3.41842  1.46634  0.26236  1.54840  0.23900
      3   1.85387  1.73195  1.88786  0.66352      3 t - - -
          1.38629  1.38629  1.38629  1.38629
          0.06777  3.41842  3.41842  1.46634  0.26236  1.54840  0.23900
      4   0.62298  2.01106  1.78035  1.82493      4 a - - -
          1.38629  1.38629  1.38629  1.38629
          0.06777  3.41842  3.41842  1.46634  0.26236  1.54840  0.23900
      5   0.62298  2.01106  1.78035  1.82493      5 a - - -
          1.38629  1.38629  1.38629  1.38629
          0.06777  3.41842  3.41842  1.46634  0.26236  1.54840  0.23900
      6   0.62298  2.01106  1.78035  1.82493      6 a - - -
          1.38629  1.38629  1.38629  1.38629
          0.06777  3.41842  3.41842  1.46634  0.26236  1.54840  0.23900
      7   1.90507  2.16817  0.51598  1.96692      7 g - - -
          1.38629  1.38629  1.38629  1.38629
          0.06777  3.41842  3.41842  1.46634  0.26236  1.54840  0.23900
      8   1.90507  2.16817  0.51598  1.96692      8 g - - -
          1.38629  1.38629  1.38629  1.38629
          0.06777  3.41842  3.41842  1.46634  0.26236  1.54840  0.23900
      9   1.85387  1.73195  1.88786  0.66352      9 t - - -
          1.38629  1.38629  1.38629  1.38629
          0.06777  3.41842  3.41842  1.46634  0.26236  1.54840  0.23900
     10   1.85387  1.73195  1.88786  0.66352     10 t - - -
          1.38629  1.38629  1.38629  1.38629
          0.06777  3.41842  3.41842  1.46634  0.26236  1.54840  0.23900
     11   1.85387  1.73195  1.88786  0.66352     11 t - - -
          1.38629  1.38629  1.38629  1.38629
          0.06777  3.41842  3.41842  1.46634  0.26236  1.54840  0.23900
     12   0.62298  2.01106  1.78035  1.82493     12 a - - -
          1.38629  1.38629  1.38629  1.38629
          0.06777  3.41842  3.41842  1.46634  0.26236  1.54840  0.23900
     13   1.85387  1.73195  1.88786  0.66352     13 t - - -
          1.38629  1.38629  1.38629  1.38629
          0.06777  3.41842  3.41842  1.46634  0.26236  1.54840  0.23900
     14   0.62298  2.01106  1.78035  1.82493     14 a - - -
          1.38629  1.38629  1.38629  1.38629
          0.06777  3.41842  3.41842  1.46634  0.26236  1.54840  0.23900
     15   1.91393  0.66933  1.96289  1.60945     15 c - - -
          1.38629  1.38629  1.38629  1.38629
          0.06777  3.41842  3.41842  1.46634  0.26236  1.54840  0.23900
     16   1.91393  0.66933  1.96289  1.60945     16 c - - -
"covid.hmm" 89732L, 5293424C
```

Figure 4: Excerpt from the resulting HMM

3 Querying the profile HMM

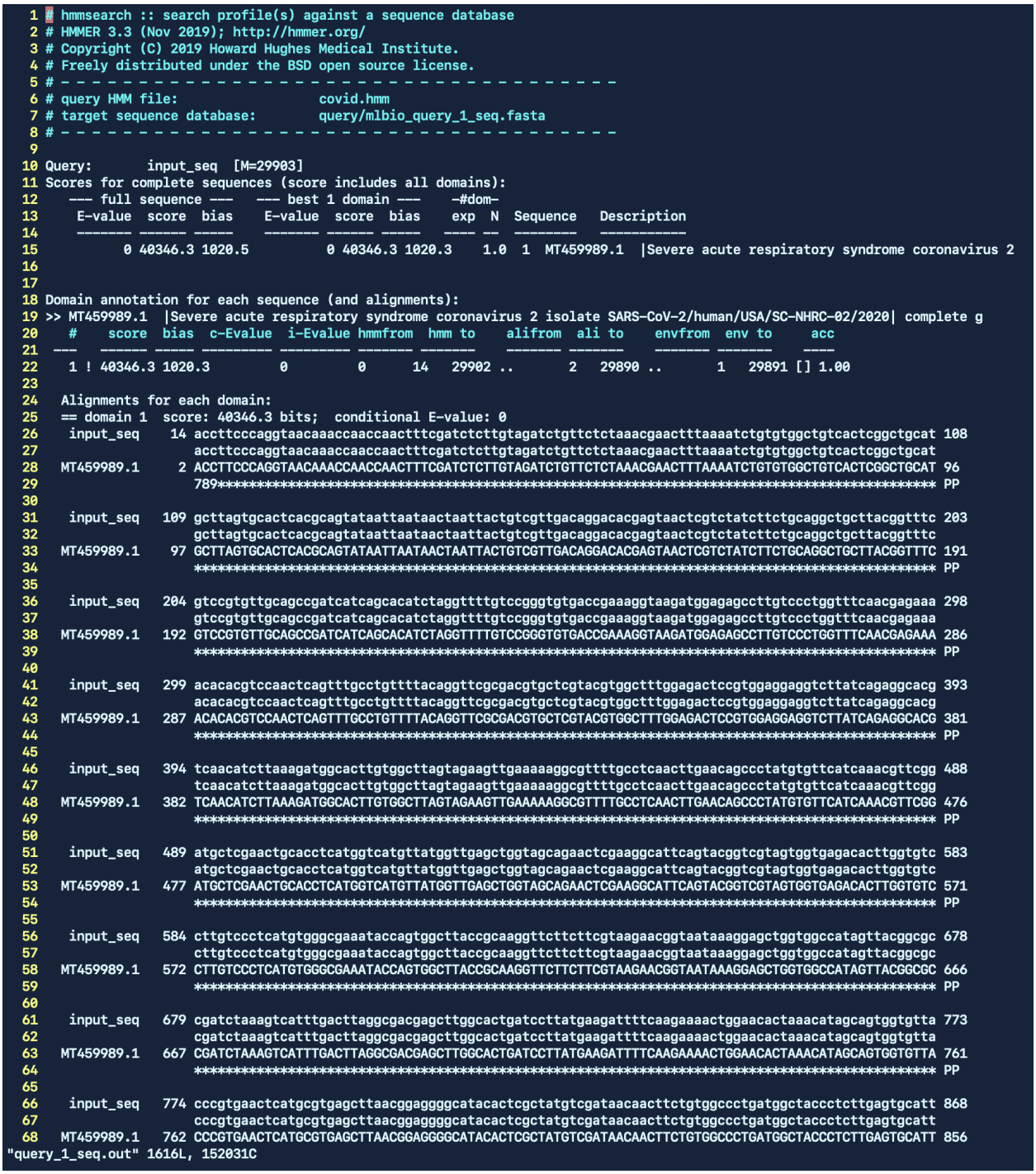
We can now use the obtained profile HMM to query it against sequence databases. In our case we have three such query databases. To run a search query of the profile HMM against the sequence database we use the `hmmsearch` command as can be seen in Figure 5. We redirect the shell output to another file to save it.

```
haean 福 ~/Documents/uni/Lecture Notes/2020 Spring/Machine Learning in Bioinformatics/HW2 ➤ bf68553|master⚡
[9066 ± : hmmsearch covid.hmm query/mlbio_query_1_seq.fasta > query_1_seq.out
```

Figure 5: Command to query the HMM with a query database

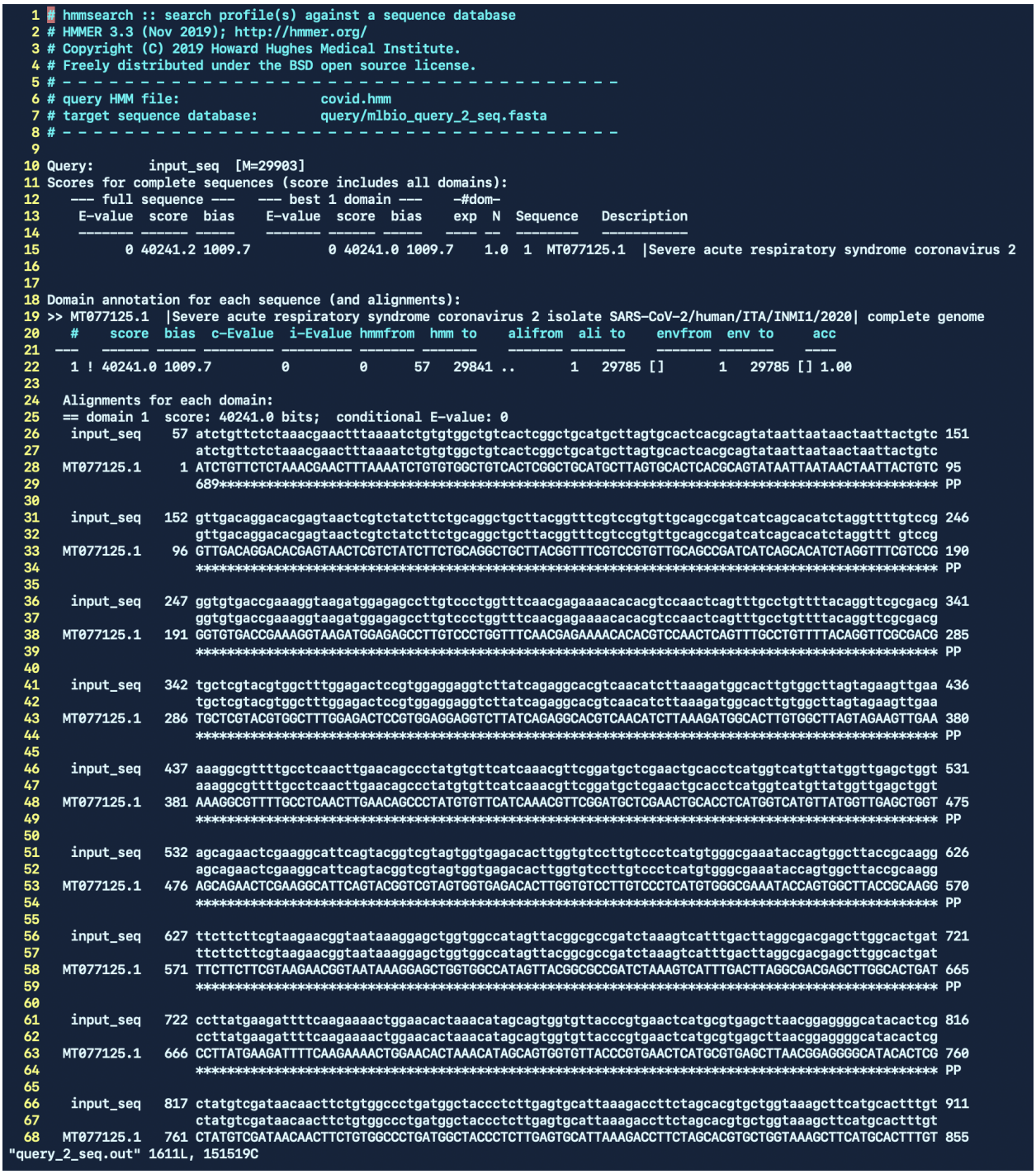
We can repeat this for each of the query databases and obtain results that can be seen in Figure 6, Figure 7 and Figure 8. Each of these sequence databases only contains a single sequence so we only see a single sequence in the resulting ranking for each of the queries (it would have been possible to query for all of these at once, see Figure 9). The most important value to look at for us is the E-value, which represents the expected number of false-positives, and thus the lower the E-value ($\ll 1$) the more likely our profile sequences are homologous to the sequence in the query database. It is thus a measure of statistical significance. We can see for all three of our queries that the E-value is 0, meaning that these sequences are highly likely to be homologous (which is not very surprising since they all are SARS-CoV-2 sequences as well). We can also look at the bit score: the higher the bit score, the more significant the query database is to our profile. For all three queries we observe a bit score $> 40\,000$, again meaning the sequences are very significant. In Figure 9 we can easily see that the sequence MT459989 is most significant as it has the highest bit score but the difference in score is very small; all of the sequences are highly significant.

Lastly, we can take a look at the "best 1 domain" scores and E-values and see that they are nearly identical to the full sequence values, meaning that our profile sequence only consists of a single domain that was found in the query sequence (if the full sequence score was high but the best single domain score was low this could mean we are dealing with a repetitive sequence).



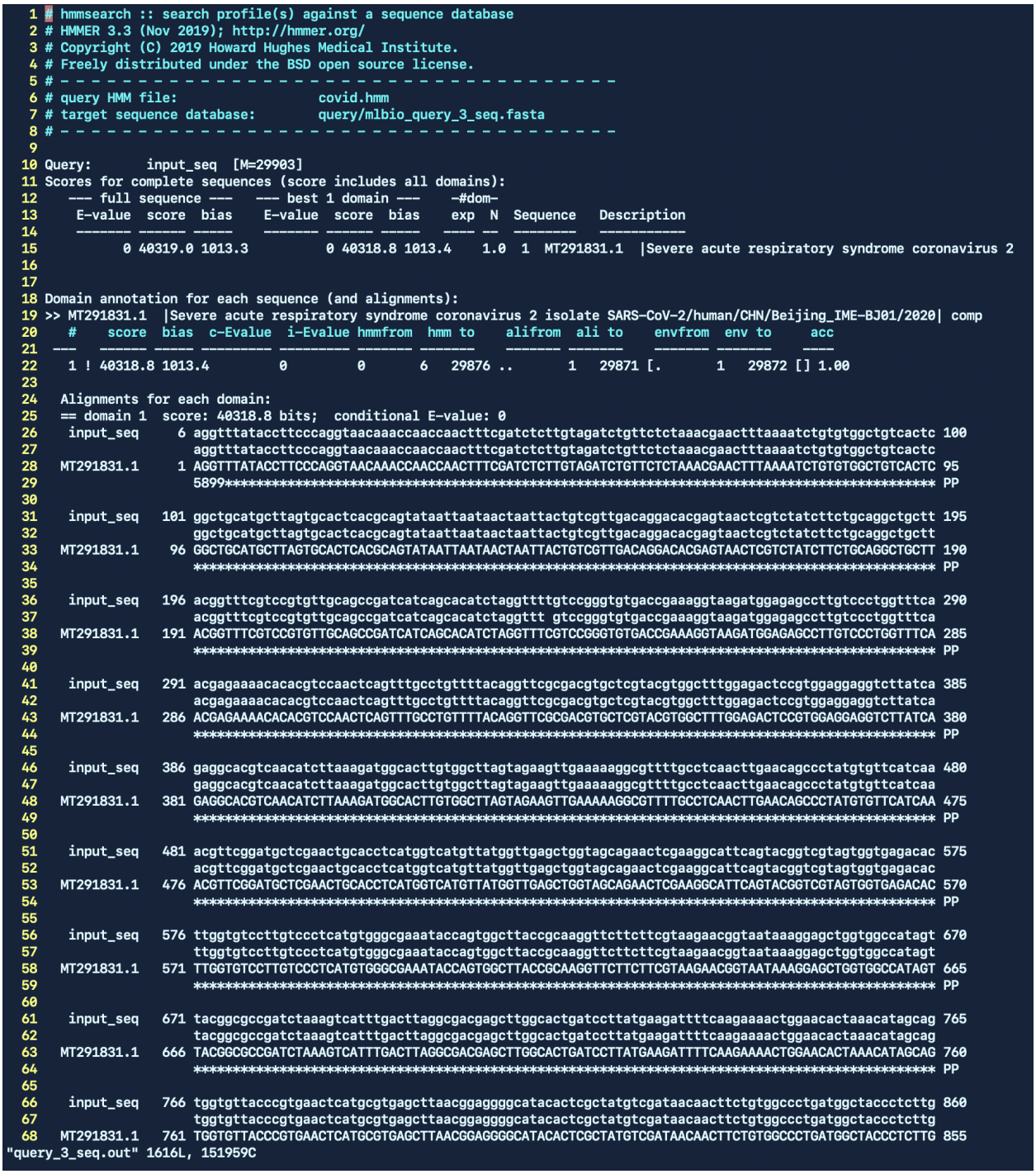
```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.3 (Nov 2019); http://hmmer.org/
# Copyright (C) 2019 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query HMM file:                  covid.hmm
# target sequence database:        query/mlbio_query_1_seq.fasta
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       input_seq  [M=29903]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence   Description
    ------- ------ -----    ------- ------ -----   ---- --  --------   -----------
          0 40346.3 1020.5          0 40346.3 1020.3    1.0  1  MT459989.1  |Severe acute respiratory syndrome coronavirus 2


Domain annotation for each sequence (and alignments):
>> MT459989.1  |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/SC-NHRC-02/2020| complete g
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 ! 40346.3 1020.3         0         0      14   29902 ..       2   29890 ..       1   29891 [] 1.00

  Alignments for each domain:
  == domain 1  score: 40346.3 bits;  conditional E-value: 0
   input_seq     14 accttcccaggtaacaaaccaaccaactttcgatctcttgtagatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcat 108
                    accttcccaggtaacaaaccaaccaactttcgatctcttgtagatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcat
   MT459989.1     2 ACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCAT 96
                    789*********************************************************************************************** PP

   input_seq    109 gcttagtgcactcacgcagtataattaataactaattactgtcgttgacaggcacgagtaactcgtctatcttctgcaggctgcttacggtttc 203
                    gcttagtgcactcacgcagtataattaataactaattactgtcgttgacaggcacgagtaactcgtctatcttctgcaggctgcttacggtttc
   MT459989.1     97 GCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGGCACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTC 191
                    *********************************************************************************************** PP

   input_seq    204 gtccgtgttgcagccgatcatcagcacatctaggttttgtccgggtgtgaccgaaaggtaagatggagagccttgtccctggtttcaacgagaaa 298
                    gtccgtgttgcagccgatcatcagcacatctaggttttgtccgggtgtgaccgaaaggtaagatggagagccttgtccctggtttcaacgagaaa
   MT459989.1   192 GTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTTGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAA 286
                    *********************************************************************************************** PP

   input_seq    299 acacacgtccaactcagtttgcctgttttacaggttcgcgacgtgctcgtacgtggctttggagactccgtggaggaggtcttatcagaggcacg 393
                    acacacgtccaactcagtttgcctgttttacaggttcgcgacgtgctcgtacgtggctttggagactccgtggaggaggtcttatcagaggcacg
   MT459989.1   287 ACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACG 381
                    *********************************************************************************************** PP

   input_seq    394 tcaacatcttaaagatggcacttgtggcttagtagaagttgaaaaaggcgttttgcctcaacttgaacagccctatgtgttcatcaaacgttcgg 488
                    tcaacatcttaaagatggcacttgtggcttagtagaagttgaaaaaggcgttttgcctcaacttgaacagccctatgtgttcatcaaacgttcgg
   MT459989.1   382 TCAACATCTTAAAGATGGCACTTGTGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGG 476
                    *********************************************************************************************** PP

   input_seq    489 atgctcgaactgcacctcatggtcatgttatggttgagctggtagcagaactcgaaggcattcagtacggtcgtagtggtgagacacttggtgtc 583
                    atgctcgaactgcacctcatggtcatgttatggttgagctggtagcagaactcgaaggcattcagtacggtcgtagtggtgagacacttggtgtc
   MT459989.1   477 ATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTC 571
                    *********************************************************************************************** PP

   input_seq    584 cttgtccctcatgtgggcgaaataccagtggcttaccgcaaggttcttcttcgtaagaacggtaataaaggagctggtggccatagttacggcgc 678
                    cttgtccctcatgtgggcgaaataccagtggcttaccgcaaggttcttcttcgtaagaacggtaataaaggagctggtggccatagttacggcgc
   MT459989.1   572 CTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGC 666
                    *********************************************************************************************** PP

   input_seq    679 cgatctaaagtcatttgacttaggcgacgagcttggcactgatccttatgaagattttcaagaaaactggaacactaaacatagcagtggtgtta 773
                    cgatctaaagtcatttgacttaggcgacgagcttggcactgatccttatgaagattttcaagaaaactggaacactaaacatagcagtggtgtta
   MT459989.1   667 CGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTA 761
                    *********************************************************************************************** PP

   input_seq    774 cccgtgaactcatgcgtgagcttaacggaggggcatacactcgctatgtcgataacaacttctgtggccctgatggctaccctcttgagtgcatt 868
                    cccgtgaactcatgcgtgagcttaacggaggggcatacactcgctatgtcgataacaacttctgtggccctgatggctaccctcttgagtgcatt
   MT459989.1   762 CCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGCCCTGATGGCTACCCTCTTGAGTGCATT 856
"query_1_seq.out" 1616L, 152031C
```

Figure 6: Result of query 1



```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.3 (Nov 2019); http://hmmer.org/
# Copyright (C) 2019 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query HMM file:                  covid.hmm
# target sequence database:        query/mlbio_query_2_seq.fasta
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       input_seq  [M=29903]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence   Description
    ------- ------ -----    ------- ------ -----   ---- --  --------   -----------
          0 40241.2 1009.7          0 40241.0 1009.7    1.0  1  MT077125.1  |Severe acute respiratory syndrome coronavirus 2


Domain annotation for each sequence (and alignments):
>> MT077125.1  |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/ITA/INMI1/2020| complete genome
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 ! 40241.0 1009.7         0         0      57   29841 ..       1   29785 []       1   29785 [] 1.00

  Alignments for each domain:
  == domain 1  score: 40241.0 bits;  conditional E-value: 0
   input_seq     57 atctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcagtataattaataactaattactgtc 151
                    atctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcagtataattaataactaattactgtc
   MT077125.1     1 ATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTC 95
                    689**********************************************************************************************  PP

   input_seq    152 gttgacaggacacgagtaactcgtctatcttctgcaggctgcttacggtttcgtccgtgttgcagccgatcatcagcacatctaggttttgtccg 246
                    gttgacaggacacgagtaactcgtctatcttctgcaggctgcttacggtttcgtccgtgttgcagccgatcatcagcacatctaggttt gtccg
   MT077125.1     96 GTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCG 190
                    **********************************************************************************************  PP

   input_seq    247 ggtgtgaccgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaactcagtttgcctgttttacaggttcgcgacg 341
                    ggtgtgaccgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaactcagtttgcctgttttacaggttcgcgacg
   MT077125.1    191 GGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACG 285
                    **********************************************************************************************  PP

   input_seq    342 tgctcgtacgtggctttggagactccgtggaggaggtcttatcagaggcacgtcaacatcttaaagatggcacttgtggcttagtagaagttgaa 436
                    tgctcgtacgtggctttggagactccgtggaggaggtcttatcagaggcacgtcaacatcttaaagatggcacttgtggcttagtagaagttgaa
   MT077125.1    286 TGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGGCTTAGTAGAAGTTGAA 380
                    **********************************************************************************************  PP

   input_seq    437 aaaggcgttttgcctcaacttgaacagccctatgtgttcatcaaacgttcggatgctcgaactgcacctcatggtcatgttatggttgagctggt 531
                    aaaggcgttttgcctcaacttgaacagccctatgtgttcatcaaacgttcggatgctcgaactgcacctcatggtcatgttatggttgagctggt
   MT077125.1    381 AAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGT 475
                    **********************************************************************************************  PP

   input_seq    532 agcagaactcgaaggcattcagtacggtcgtagtggtgagacacttggtgtccttgtccctcatgtgggcgaaataccagtggcttaccgcaagg 626
                    agcagaactcgaaggcattcagtacggtcgtagtggtgagacacttggtgtccttgtccctcatgtgggcgaaataccagtggcttaccgcaagg
   MT077125.1    476 AGCAGAACTCGAAGGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGG 570
                    **********************************************************************************************  PP

   input_seq    627 ttcttcttcgtaagaacggtaataaaggagctggtggccatagttacggcgccgatctaaagtcatttgacttaggcgacgagcttggcactgat 721
                    ttcttcttcgtaagaacggtaataaaggagctggtggccatagttacggcgccgatctaaagtcatttgacttaggcgacgagcttggcactgat
   MT077125.1    571 TTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGAT 665
                    **********************************************************************************************  PP

   input_seq    722 ccttatgaagattttcaagaaaactggaacactaaacatagcagtggtgttacccgtgaactcatgcgtgagcttaacggaggggcatacactcg 816
                    ccttatgaagattttcaagaaaactggaacactaaacatagcagtggtgttacccgtgaactcatgcgtgagcttaacggaggggcatacactcg
   MT077125.1    666 CCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCG 760
                    **********************************************************************************************  PP

   input_seq    817 ctatgtcgataacaacttctgtggccctgatggctaccctcttgagtgcattaaagaccttctagcacgtgctggtaaagcttcatgcactttgt 911
                    ctatgtcgataacaacttctgtggccctgatggctaccctcttgagtgcattaaagaccttctagcacgtgctggtaaagcttcatgcactttgt
   MT077125.1    761 CTATGTCGATAACAACTTCTGTGGCCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTGT 855
"query_2_seq.out" 1611L, 151519C
```

Figure 7: Result of query 2



```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.3 (Nov 2019); http://hmmer.org/
# Copyright (C) 2019 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query HMM file:                  covid.hmm
# target sequence database:        query/mlbio_query_3_seq.fasta
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       input_seq  [M=29903]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence   Description
    ------- ------ -----    ------- ------ -----   ---- --  --------   -----------
          0 40319.0 1013.3          0 40318.8 1013.4    1.0  1  MT291831.1  |Severe acute respiratory syndrome coronavirus 2


Domain annotation for each sequence (and alignments):
>> MT291831.1  |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/CHN/Beijing_IME-BJ01/2020| comp
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 ! 40318.8 1013.4         0         0       6   29876 ..       1   29871 [.       1   29872 [] 1.00

  Alignments for each domain:
  == domain 1  score: 40318.8 bits;  conditional E-value: 0
   input_seq     6 aggtttatacctttcccaggtaacaaaccaaccaactttcgatctcttgtagatctgttctctaaacgaactttaaaatctgtgtggctgtcactc 100
                   aggtttatacctttcccaggtaacaaaccaaccaactttcgatctcttgtagatctgttctctaaacgaactttaaaatctgtgtggctgtcactc
   MT291831.1     1 AGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTC 95
                   5899**********************************************************************************************  PP

   input_seq   101 ggctgcatgcttagtgcactcacgcagtataattaataactaattactgtcgttgacaggacacgagtaactcgtctatcttctgcaggctgctt 195
                   ggctgcatgcttagtgcactcacgcagtataattaataactaattactgtcgttgacaggacacgagtaactcgtctatcttctgcaggctgctt
   MT291831.1    96 GGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTT 190
                   **********************************************************************************************  PP

   input_seq   196 acggtttcgtccgtgttgcagccgatcatcagcacatctaggttttgtccgggtgtgaccgaaaggtaagatggagagccttgtccctggtttca 290
                   acggtttcgtccgtgttgcagccgatcatcagcacatctaggttt gtccgggtgtgaccgaaaggtaagatggagagccttgtccctggtttca
   MT291831.1    191 ACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCA 285
                   **********************************************************************************************  PP

   input_seq   291 acgagaaaacacacgtccaactcagtttgcctgttttacaggttcgcgacgtgctcgtacgtggctttggagactccgtggaggaggtcttatca 385
                   acgagaaaacacacgtccaactcagtttgcctgttttacaggttcgcgacgtgctcgtacgtggctttggagactccgtggaggaggtcttatca
   MT291831.1    286 ACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCA 380
                   **********************************************************************************************  PP

   input_seq   386 gaggcacgtcaacatcttaaagatggcacttgtggcttagtagaagttgaaaaaggcgttttgcctcaacttgaacagccctatgtgttcatcaa 480
                   gaggcacgtcaacatcttaaagatggcacttgtggcttagtagaagttgaaaaaggcgttttgcctcaacttgaacagccctatgtgttcatcaa
   MT291831.1    381 GAGGCACGTCAACATCTTAAAGATGGCACTTGTGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAA 475
                   **********************************************************************************************  PP

   input_seq   481 acgttcggatgctcgaactgcacctcatggtcatgttatggttgagctggtagcagaactcgaaggcattcagtacggtcgtagtggtgagacac 575
                   acgttcggatgctcgaactgcacctcatggtcatgttatggttgagctggtagcagaactcgaaggcattcagtacggtcgtagtggtgagacac
   MT291831.1    476 ACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTCGTAGTGGTGAGACAC 570
                   **********************************************************************************************  PP

   input_seq   576 ttggtgtccttgtccctcatgtgggcgaaataccagtggcttaccgcaaggttcttcttcgtaagaacggtaataaaggagctggtggccatagt 670
                   ttggtgtccttgtccctcatgtgggcgaaataccagtggcttaccgcaaggttcttcttcgtaagaacggtaataaaggagctggtggccatagt
   MT291831.1    571 TTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGT 665
                   **********************************************************************************************  PP

   input_seq   671 tacggcgccgatctaaagtcatttgacttaggcgacgagcttggcactgatccttatgaagattttcaagaaaactggaacactaaacatagcag 765
                   tacggcgccgatctaaagtcatttgacttaggcgacgagcttggcactgatccttatgaagattttcaagaaaactggaacactaaacatagcag
   MT291831.1    666 TACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAG 760
                   **********************************************************************************************  PP

   input_seq   766 tggtgttacccgtgaactcatgcgtgagcttaacggaggggcatacactcgctatgtcgataacaacttctgtggccctgatggctaccctcttg 860
                   tggtgttacccgtgaactcatgcgtgagcttaacggaggggcatacactcgctatgtcgataacaacttctgtggccctgatggctaccctcttg
   MT291831.1    761 TGGTGTTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGCCCTGATGGCTACCCTCTTG 855
"query_3_seq.out" 1616L, 151959C
```

Figure 8: Result of query 3

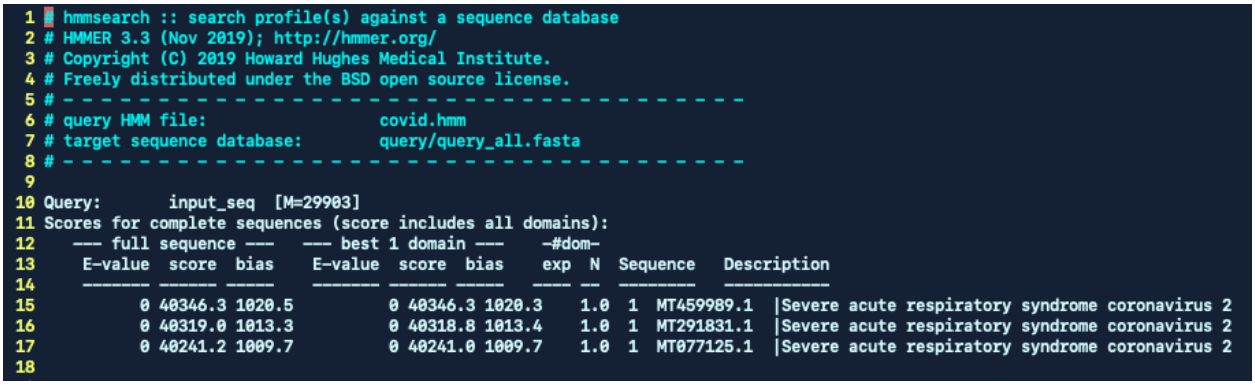


Figure 9: Results when querying all databases at once

References

[1] Edgar, Robert C. "MUSCLE: a multiple sequence alignment method with reduced time and space complexity." *BMC Bioinformatics* 5.1 (2004): 113.

[2] Eddy, Sean. "HMMER user's guide." *Department of Genetics, Washington University School of Medicine* 2.1 (1992): 13. See: