Homework 2

Björn Bebensee (2019–21343)

Machine Learning in Bioinformatics

June 10, 2020

# 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.

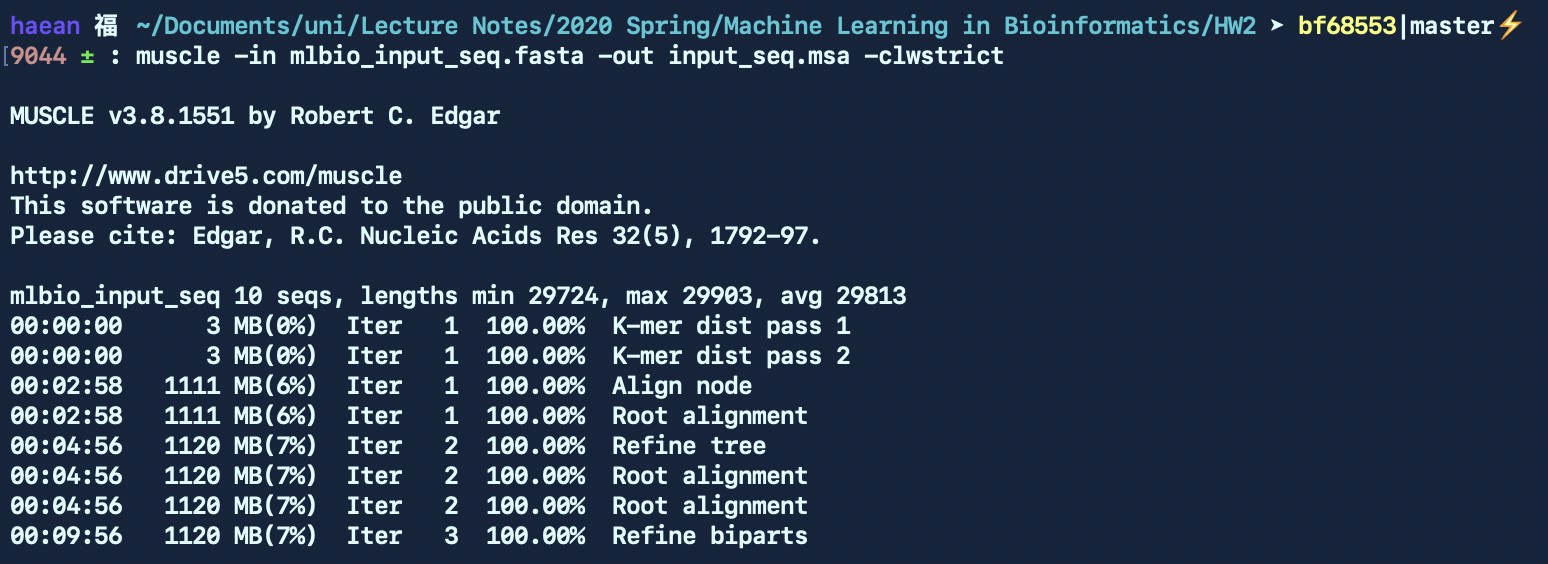


Figure 1: Command used to align the sequences

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.

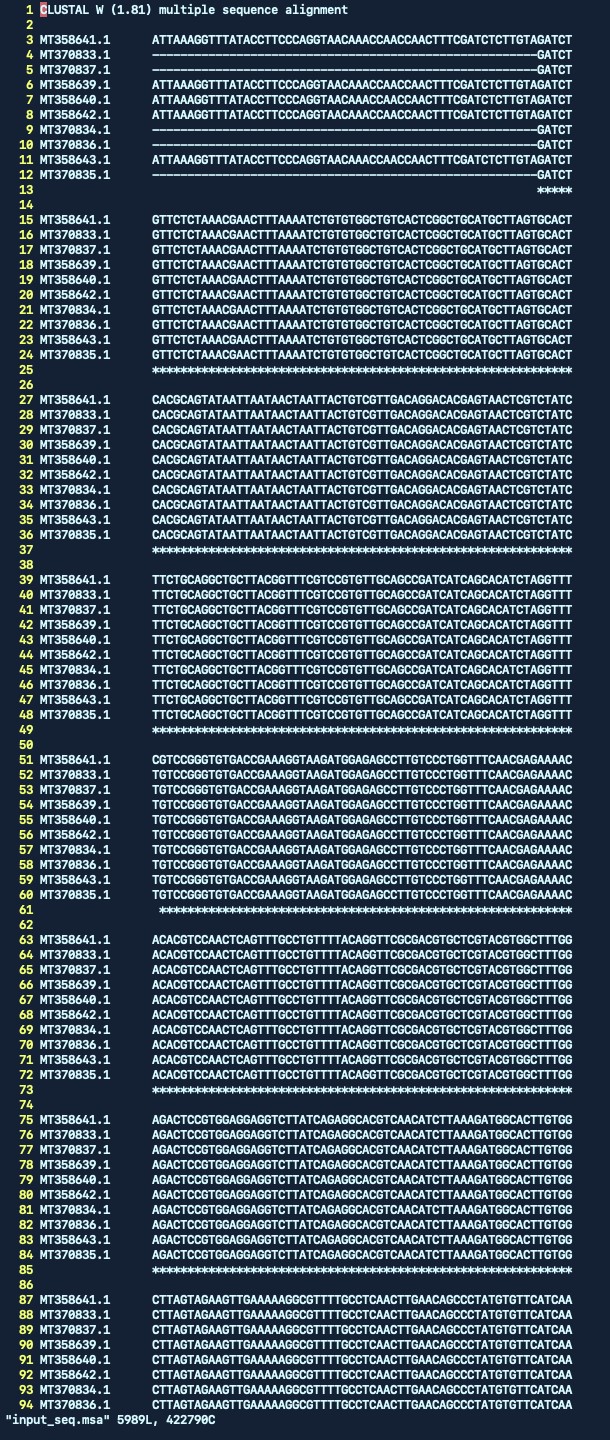


Figure 2: Resulting multiple sequence alignment

# 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 29903 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).

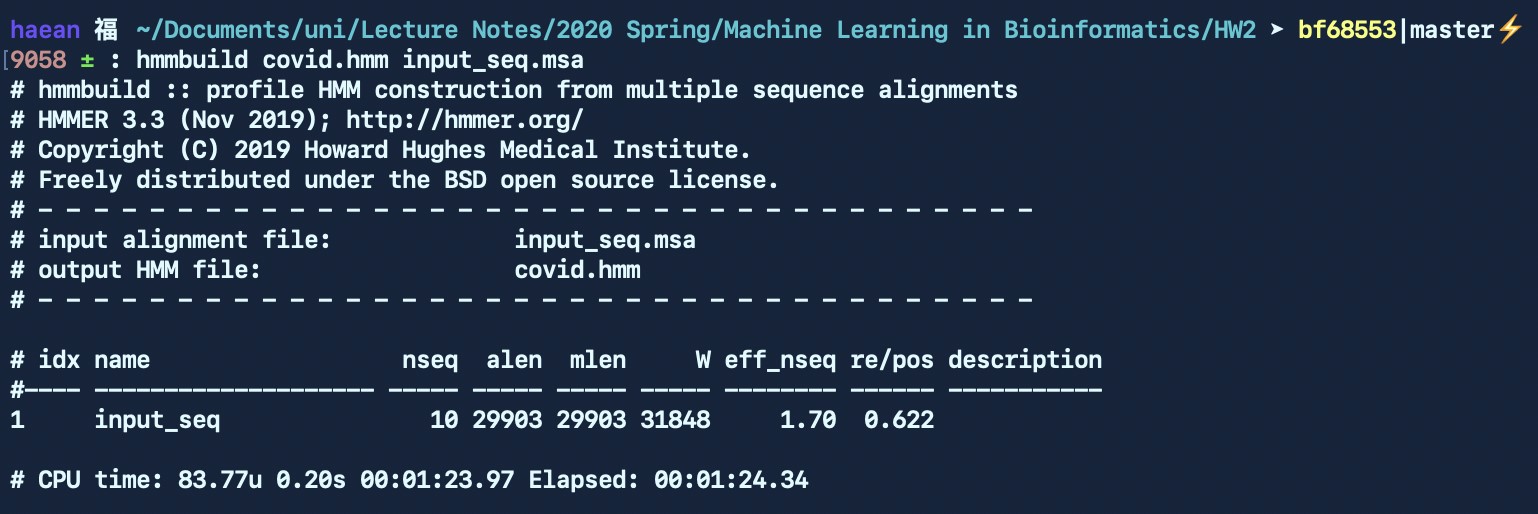


Figure 3: Command to build the HMM

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 29903) contains three lines: match emission probabilities, insert emission probabilities and finally transition probabilities between the states.

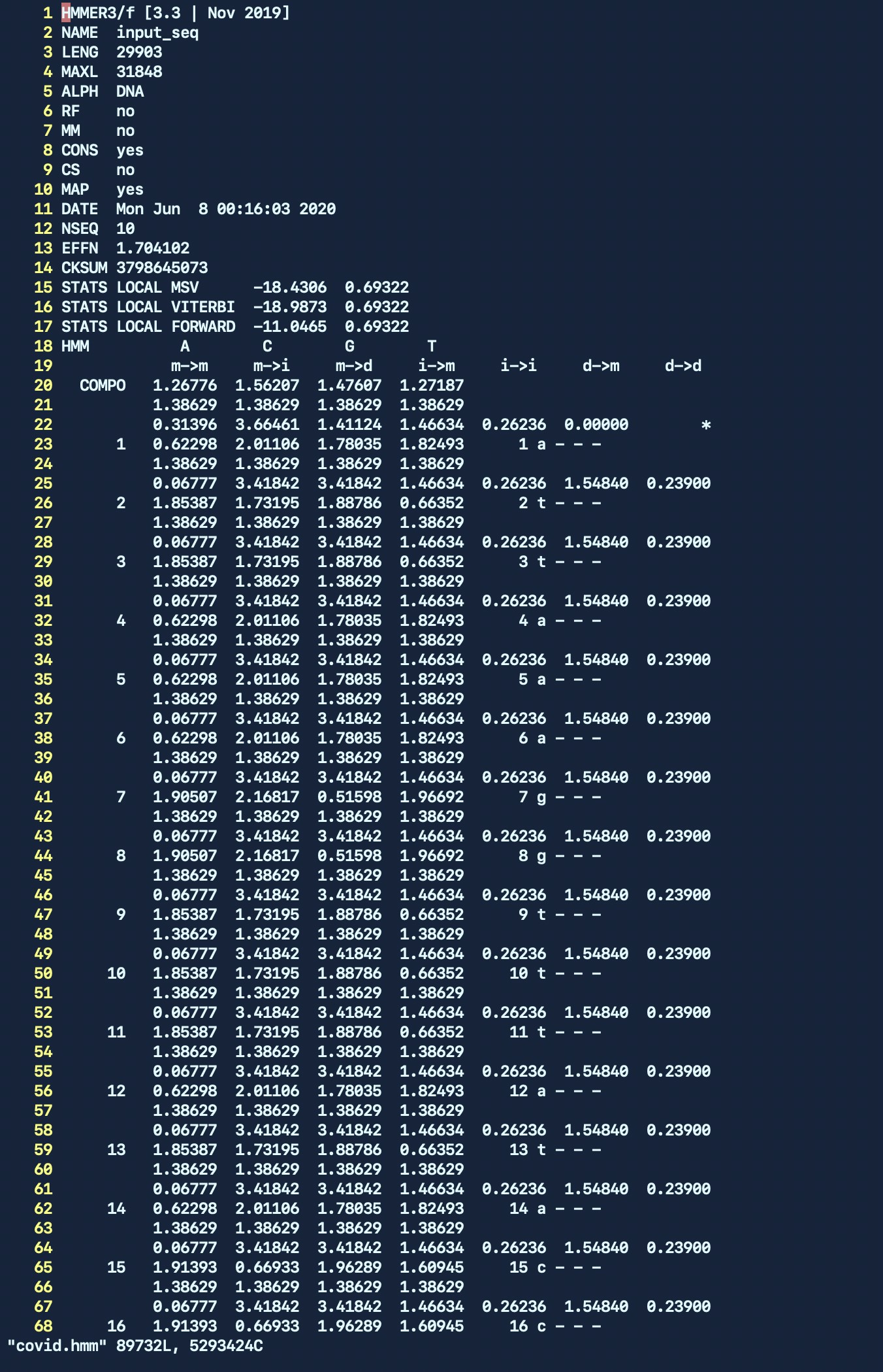


Figure 4: Excerpt from the resulting HMM

# 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.



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 (<<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 *>* 40000, 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).

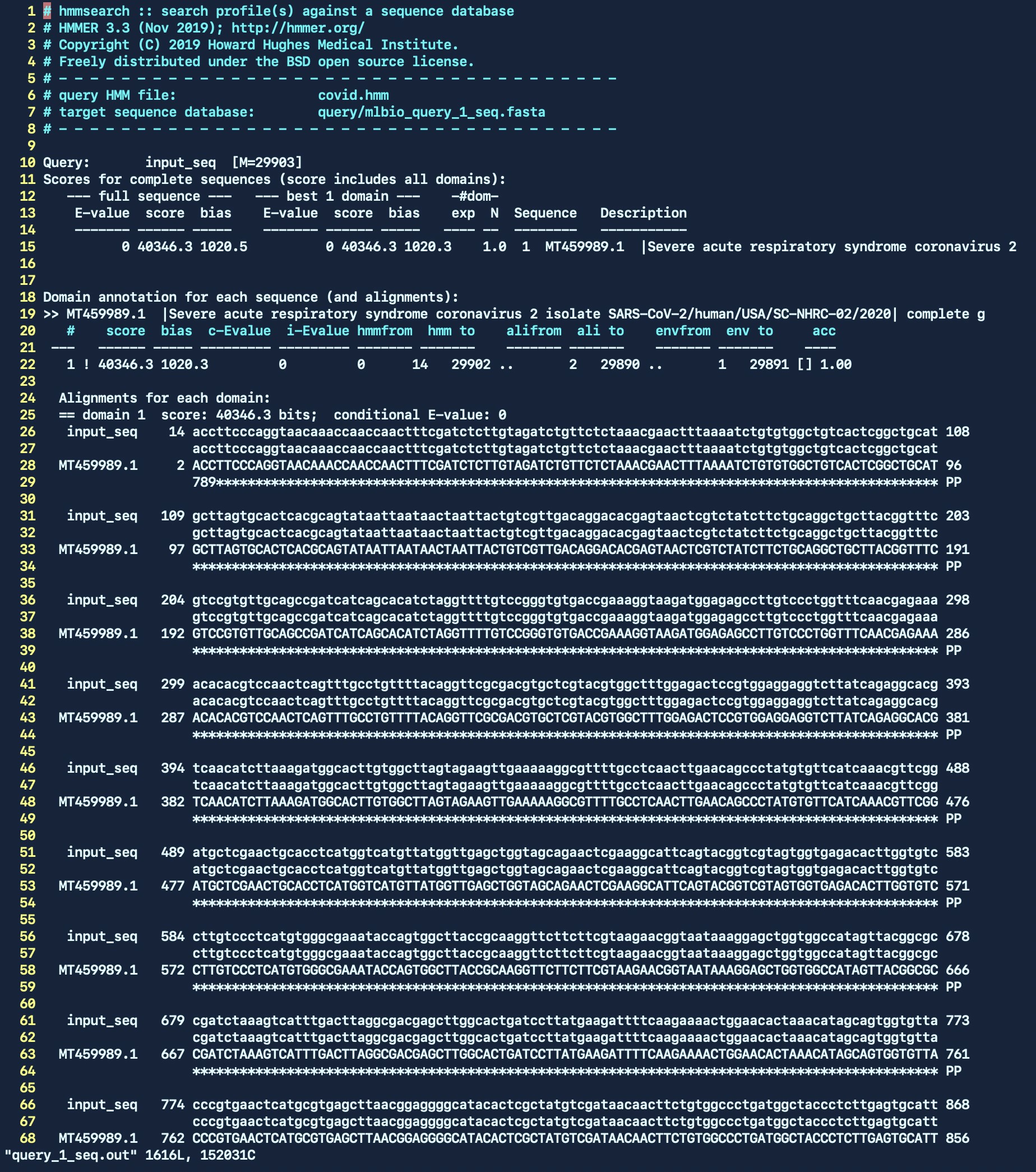


Figure 6: Result of query 1

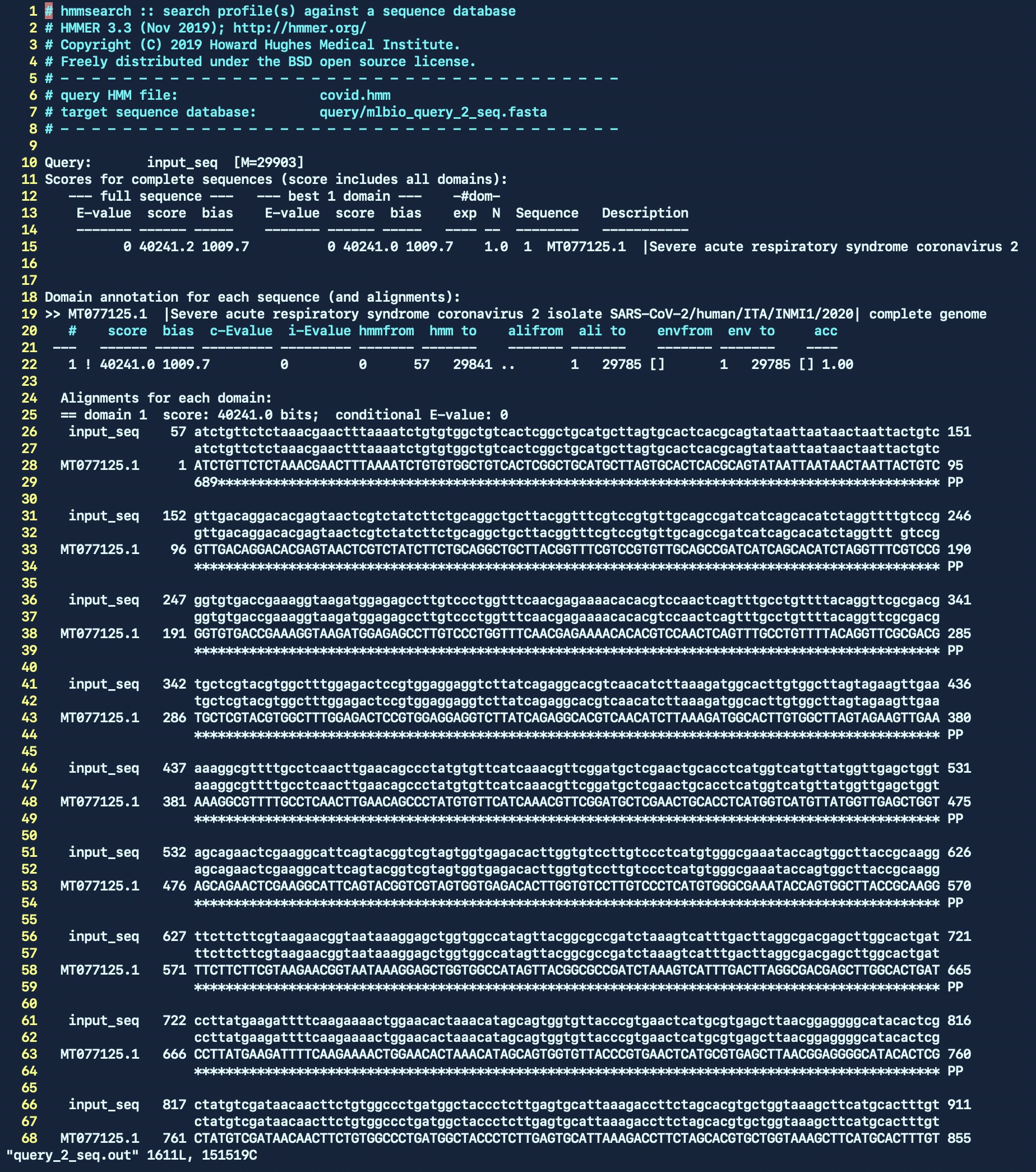


Figure 7: Result of query 2

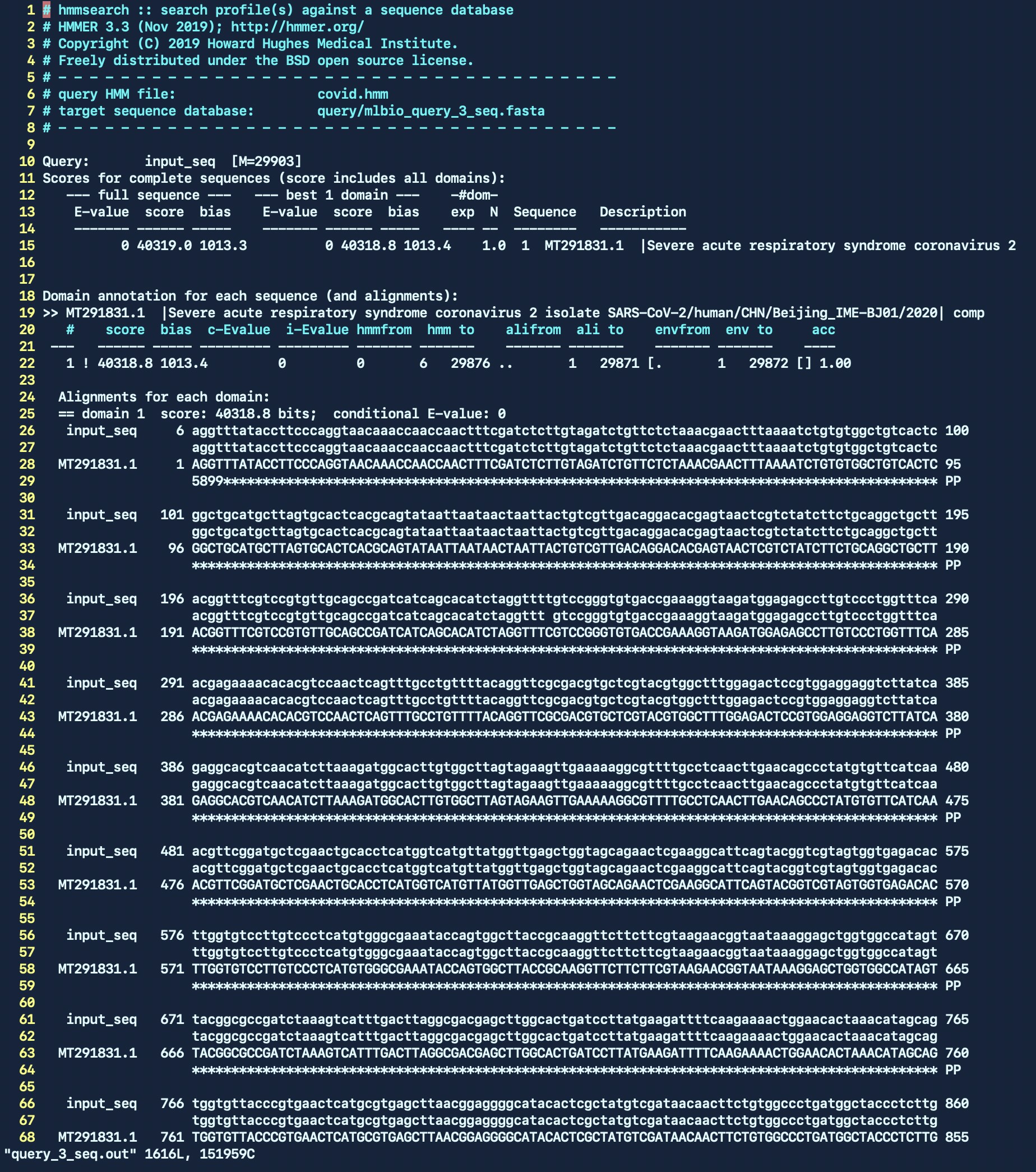


Figure 8: Result of query 3

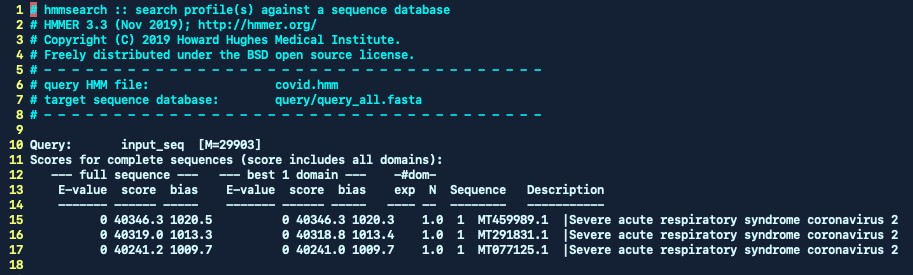


Figure 9: Results when querying all databases at once

# 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.”](http://eddylab.org/software/hmmer/Userguide.pdf) *Department of Genetics, Washington University School of Medicine* 2.1 (1992): 13. See: