Data Science and Machine Learning for The Biosciences: Code Report

GitHub file:

In December 2019, the first case of an unknown respiratory disease was reported in Wuhan, China (Wu *et al.* 2020). Now, over a year later in January 2021, this unknown respiratory disease is known as COVID-19 and to date, has been responsible for over 1.8 million deaths worldwide (Dong *et al.* 2020). Metagenomic RNA sequencing was undertaken from bronchoalveolar lavage fluid of a patient identified with COVID-19 (Wu *et al.* 2020). A viral genome was then assembled from this data which revealed that COVID-19 was caused by a new strain of virus never seen before, called SARS-CoV-2.

This Jupyter Notebook develops previous work presented to explore how languages such as Python can be utilised to analyse and interpret different genomic sequences, providing better insights into experimental data obtained. Ultimately, demonstrating the utility of such computer languages such as Python to provide data driven insights into potential therapeutic targets for the disease.

In this analysis, the SARS-CoV-2 (COVID-19) sequence was analysed. The number of each respective nucleotide base was counted, producing the results: A: 8954, T: 9594, G: 5863, C: 5492. The GC% was then calculated producing as result of 38%. It can be useful to know the GC% of a sequence when designing primers for experimental analysis (Lorenz, 2012). The sequence was then translated and transcribed into its amino acids and specific protein chains were identified where stop codons were present, represented by a ‘\*’.

A key disadvantage to this code is that it does not consider modifications that are made to generate mature mRNA or post-translational modifications made to proteins. Therefore, when translating the mRNA into its amino acids, a warning message occurs, as 29,903 is not a multiple of 3 which is required to accurately translate mRNA. For example, the poly(A) is removed and not translated into protein (Rubin and Halim, 1992). However, this was not included in the code as this could be one of many modifications. This code also does not consider start codons, usually AUG, which translates into the amino acid methionine (Firth and Brierley, 2012). This was not included as a method to split protein sequences as it does not consider non-canonical initiation of translation where AUG is not required.

Protein sequences less than 20 amino acids were removed as the smallest functional known protein is 20 amino acids long (Neidigh *et al.* 2002). These were then collated into a dataframe. Individual files were also created for each protein sequence identified. This Jupyter Notebook also provides code for calculating certain properties of each suggested protein sequence, including: Amino Acids Percent, Molecular Weight, Aromaticity, Flexibility, Isoelectric Point and Secondary Structure Fraction. This is an important step in analysing and beginning to understand how a protein may function in biological systems.

The longest protein sequence identified, Peptide\_49.fasta, was chosen for further investigation as it was considerably larger than the remaining discovered protein sequences. Peptide\_49.fasta was then searched against NCBI's Protein Blast database within Python. Other sequences could also be searched for using this code by supplying a different fasta file in the code. This matched to the ORF1ab polyprotein in SARS-CoV-2 (COVID-19), with a summary of the results written and saved to a separate file called Protein49out.xml. The ORF1ab polyprotein is one of the largest and complex polyproteins which encodes for a set of polyproteins which form a replicase-transcriptase complex important for RNA synthesis of the virus (Banerjee *et al.* 2020; Gordon *et al.* 2020). This highlights a potential therapeutic target in treating COVID-19 and the benefits of using Python and associated packages to analyse genomic sequences.

This Jupyter Notebook also provides code for primer design using Primer3. Here, primers were designed for the ORF1ab polyprotein identified. Three separate regions within this protein were designated for primer design, with five separate primer pairs being designed for each specified region. The code used here can also be modified to include several different sequences or different regions of a particular sequence for primer design. A disadvantage of this code is that you need to know the specific region from a nucleotide sequence that you want to design primers for as this is not included in the code. In the case of ORF1ab polyprotein, NCBI was used to determine the specific region required for primer design.

Finally, code is provided to compare the similarity in nucleotide sequence between SARS-CoV-2 (COVID-19) and both Severe Acute Respiratory Syndrome (SARS-CoV-1) and Middle East Respiratory Syndrome (MERS-CoV). This found that SARS-CoV-1 is most like COVID-19, sharing a similarity of 83.3%, whilst MERS-CoV shared a similarity of 69.4%. The code here can also be modified and used to compare other sequences of interest.

In conclusion, this Jupyter Notebook has been developed to interpret and analyse the SARS-CoV-2 (COVID-19) sequence. However, any fasta sequence could be interrogated here. The information gathered here can be utilised, in conjunction with experimental data, to gain further understanding about a novel virus which has halted the world. Ultimately, providing key insights into potential therapeutic targets. This Jupyter Notebook is a prime example for how coding languages, such as Python, can be utilised to gleam further information from sequencing results.

References

Banerjee, S., Seal, S., Dey, R., Mondal, K.Kr. and Bhattacharjee, P. 2020. Mutational spectra of SARS-CoV-2 orf1ab polyprotein and signature mutations in the United States of America. *Journal of Medical Virology* n/a(n/a). Available at: <https://doi.org/10.1002/jmv.26417> [Accessed: 5 January 2021].

Dong, E., Du, H. and Gardner, L. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* 20(5), pp. 533–534. doi: [10.1016/S1473-3099(20)30120-1](https://doi.org/10.1016/S1473-3099(20)30120-1).

Firth, A.E. and Brierley, I. 2012. Non-canonical translation in RNA viruses. *The Journal of general virology* 93(Pt 7), pp. 1385–1409. doi: [10.1099/vir.0.042499-0](https://doi.org/10.1099/vir.0.042499-0).

Gordon, D.E. et al. 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583(7816), pp. 459–468. doi: [10.1038/s41586-020-2286-9](https://doi.org/10.1038/s41586-020-2286-9).

Lorenz, T.C. 2012. Polymerase chain reaction: basic protocol plus troubleshooting and optimization strategies. *Journal of visualized experiments : JoVE* (63), pp. e3998–e3998. doi: [10.3791/3998](https://doi.org/10.3791/3998).

Neidigh, J.W., Fesinmeyer, R.M. and Andersen, N.H. 2002. Designing a 20-residue protein. *Nature Structural Biology* 9(6), pp. 425–430. doi: [10.1038/nsb798](https://doi.org/10.1038/nsb798).

Rubin, H.N. and Halim, M.N. 1993. Why, when and how does the poly(A) tail shorten during mRNA translation? *International Journal of Biochemistry* 25(3), pp. 287–295. doi: [10.1016/0020-711X(93)90615-L](https://doi.org/10.1016/0020-711X(93)90615-L).

Wu, F. et al. 2020. A new coronavirus associated with human respiratory disease in China. *Nature* 579(7798), pp. 265–269. doi: [10.1038/s41586-020-2008-3](https://doi.org/10.1038/s41586-020-2008-3).