Detecting Selection from Longitudinal Deep Sequencing Data in Viruses

Bhavin Khatri
bkhatri@imperial.ac.uk

in collaboration with Matteo Fumagalli
m.fumagalli@imperial.ac.uk

Talia Al-Mushadani
ta1915@ic.ac.uk

1 Keywords

Evolutionary Virology, Population Genetics, Natural Selection, Machine Learning, Norovirus

2 Introduction to Project & Proposed Questions

Currently research into how the Norovirus genome evolves via selection has been limited in scope (Steyer et al., 2018) due to the lack of longitudinal deep-sequencing data on the virus and the lack of a simple methods to assess where in the genome selection is occurring. This project will be using longitudinal deep-sequencing data of Norovirus in patients, provided by Prof. Judith Breuer at UCL/GOSH, and will compare two streamlined identification methodologies. The two main methods that will be compared are a heuristic approach (Khatri, 2016) and neural networks, thereby providing a contrast between methods built on theory vs methods which are biologically uninformed. As such this project has two main questions that it wishes to address:
1) Which method is most effective at accurately predicting selection?
2) Which loci are under selection in the Norovirus genome in each individual?

3 Proposed Methods

Initially the heuristic approach algorithm (Khatri, 2016) will be tested using simulation longitudinal deep-sequencing data for a single loci with a known selection strength, employing a multiple hypothesis procedure with false discovery rates (Benjamini and Hochberg, 1995), when identifying whether the peaks in nucleotide frequency in the loci over time are characteristic of selection. A recurrent neural network (RNN) will then be created (using the same simulated data) to be able to recognise the strength of selection in the single loci.
Both of these methods will then be extended to be able to look at 2 loci. Finally these methods would be expanded to take into account an entire section of viral DNA, which would involve producing another neural network consisting, possibly, of exchangeable convolutional layers (Chan et al., 2018)

with Markov intuition, that are capable of using image data for identifying selection.

The accuracy of the methods at each stage will be assessed using simulated data, before applying the most accurate methods to the real Norovirus patient data, to identify which specific loci are under selection.

If time permits, an additional numerical path integrating approach may also be considered, which is based upon theory commonly found in computational physics (the method is outlined in an as yet unpublished paper by B. Khatri).

4 Anticipated Outcomes

I anticipate that at the end of this project I will have been able to identify loci in the Norovirus genomes from each patient which are under selection. This would enable further research into the identification of the specific genes that are being selected to evolve whilst Norovirus is infecting a host, allowing for the generation of new hypotheses which can be investigated via molecular assays.

5 Project Feasibility & Proposed Timeline

The project is feasible as the heuristic approach has already been shown to work on single loci simulations (Khatri, 2016), and the theory behind the algorithm is applicable to multiple loci. Also, most of the machine learning techniques that will be employed have already been applied to longitudinal data (Choi et al., 2017; Wang et al., 2018), although some changes to the current exchangeable neural network structure (Chan et al., 2018) would need to be made for the purposes of this project. This however is all dependent on the amount of time taken by each step, as to the number of loci that can be considered at once.

6 Itemised Budget

Research-related travel: £420

As I do not live at Silwood, and will be occupying desk space at Silwood

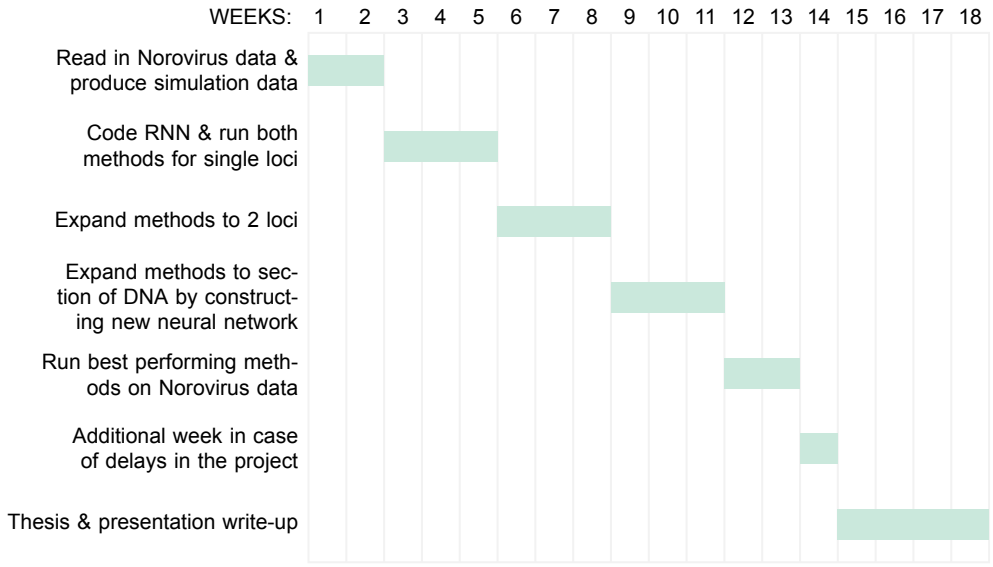

Figure 1: Gantt chart for this project

instead of at the Crick Institute, where my main supervisor is mainly based, an outlay of £420, which accounts to 35 days return train fare (half the number of days that I expect to travel into Silwood over the course of 14 weeks) would be of great assistance to me.

7 Cited References

Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing'.
URL: *https://www.jstor.org/stable/2346101*

Chan, J., Perrone, V., Spence, J. P., Jenkins, P. A., Mathieson, S. and Song, Y. S. (2018), A Likelihood-Free Inference Framework for Population Genetic Data using Exchangeable Neural Networks, Technical report.
URL: *https://github.com/popgenmethods/*

Choi, E., Schuetz, A., Stewart, W. F. and Sun, J. (2017), 'Using recurrent neural network models for early detection of heart failure onset.', *Journal of the American Medical Informatics Association : JAMIA* **24**(2), 361–370.

URL: *http://www.ncbi.nlm.nih.gov/pubmed/27521897*
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5391725

Khatri, B. S. (2016), 'Quantifying evolutionary dynamics from variant-frequency time series', *Scientific Reports* **6**(1), 32497.
URL: *http://www.nature.com/articles/srep32497*

Steyer, A., Konte, T., Sagadin, M., Kolenc, M., Škoberne, A., Germ, J., Dovč-Drnovšek, T., Arnol, M. and Poljšak-Prijatelj, M. (2018), 'Intrahost Norovirus Evolution in Chronic Infection Over 5 Years of Shedding in a Kidney Transplant Recipient', *Frontiers in Microbiology* **9**, 371.
URL: *http://journal.frontiersin.org/article/10.3389/fmicb.2018.00371/full*

Wang, T., Qiu, R. G. and Yu, M. (2018), 'Predictive Modeling of the Progression of Alzheimer's Disease with Recurrent Neural Networks', *Scientific Reports* **8**(1), 9161.
URL: *http://www.nature.com/articles/s41598-018-27337-w*