

Tracing Footprints of the Human Immune System in Virus Genomes with Machine Learning

Chae-Ryeong Yeo - 2458600y

September 22, 2023

1 Status report

1.1 Achievement

- Created a directory structure for project and stored in version control on remote private GitHub repository.
- Created research proposal and project timeline.
- Read through 4 relevant research papers and briefly summarised each.
- Stated the Bioinformatics Specialisation course in Coursera to familiarise with bioinformatics and genomic concepts.
- Came up with 4 different machine learning ways to approach the project, 3 of them being a classification problem and 1 being a regression problem approach.

1.2 Questions

- Meaning of topic: uncover subtle/hidden clues or indications of the human immune system's influence on virus genomes? Similarities of human T cell genome and virus genome? Must it be between the human immune system cell genome and a virus genome where the virus has attacked that human before?
- Are supervised or unsupervised learning methods appropriate? (Or should I use reinforcement learning to maximize reward signals? What will this reward signal be for?) I think unsupervised learning is most appropriate because we don't know the similarities between human immune system and virus genome. However, for some similar proteins found for both virus genome and human immune cells, I can use supervised learning - however, for what goal? For classification - whether it belongs to a virus or human? I think this is not the goal of this project. But I'm not sure what the goal is for the machine learning algorithm. Must it first know which are the footprints of the human immune system in the virus genome and try to find the similar parts to that in the virus genome for as many viruses as possible?
- Am I only concerned about the impact of the human immune system on the virus genome, and not the other way around?

- Is my project a research-focused project? Or is it a software focused project? Is my goal to produce an end to end learning system?

1.3 Plan for rest of week 1, and week 2

- Create a directory just for relevant research papers and its summaries, and read up on at least 10 more related research papers, summarise the findings and note down which parts of it can also be implemented for my project.
- Familiarise with bioinformatics concepts - continue the Bioinformatics Specialisation course in Coursera and always keep track of relevant findings.
- Write the minute on Moodle and review afterwards.
- Practice implementing different machine learning models for the first time in Python and Jupyter Notebook.