Bioinformatics Summer School Report

Name: Rakan Alsallum
Internship Organization: The Donnelly Centre for Cellular & Biomolecular Research, University of Toronto, Toronto, Canada
Supervisor: Artem Babaian

Project Title: The Search for Novel DNA Viruses in Alzheimer's Disease Brains

1. Introduction

Alzheimer's disease (AD), the leading cause of dementia, remains a significant challenge in medical research due to its unknown cause and lack of effective treatments. Recent studies increasingly suggest that neuroinflammation plays a role in the pathogenesis of AD, potentially linked to viral infections such as herpesviruses and neuroCOVID-19. Despite these findings, the involvement of infectious agents in AD is not yet fully understood.

This project seeks to investigate the potential involvement of small, single-stranded DNA viruses from the *Cirlivirales* order, particularly within the *Circoviridae* family, in the development of AD. By constructing a comprehensive database of *Cirlivirales* sequences and employing bioinformatics tools, this research aims to explore these potential connections and address a significant gap in understanding the viral contributions to AD.

2. Materials and Methodology

The objective of this research is to detect *Cirlivirales* viruses in patients sequencing data with high sensitivity and specificity, using a computational approach. To achieve this, I undertook the following steps:

- **Data Collection:** I retrieved all *Cirlivirales* capsid protein sequences from the National Center for Biotechnology Information (NCBI) RefSeq database, based on the annotated taxonomy of the sequences. Multiple Sequence Comparison by Log-Expectation was used for Multiple Sequence Alignment (MSA) due to its accuracy and efficiency in handling large datasets.
- **Database Construction:** Following data collection, I proceeded with database construction, querying all *Cirlivirales* capsid proteins to capture the full set of previously known *Circoviridae*, as obtained from Pfam. The sequences were then clustered to reduce redundancy, ensuring a more accurate and comprehensive database.
- **Structural Analysis:** I performed structural analysis to identify the jellyroll hallmark structure, a highly conserved fold common in many icosahedral viral capsids. This jellyroll structure serves as the primary indicator for identifying potential novel

Circoviridae. Molecular structure prediction tools, such as AlphaFold3, were employed to verify the presence of this structure in the sequences.

- **Hidden Markov Model:** To further improve detection capabilities, I employed HMMER, which uses Hidden Markov Models (HMMs) to identify sequence homologs. I developed specific HMM models based on the conserved jellyroll structure, essential for accurately detecting potential novel *Circoviridae* within large sequencing datasets.

This approach will enable the identification of novel viruses belonging to the *Circoviridae* family and facilitate the discovery of new viral species and strains by offering a robust comparative framework.

3. Results and Discussion

The project successfully created a comprehensive *Cirlivirales* jellyroll database, validated for the presence of the jellyroll structure. Although screening of sequencing data has not yet begun, this database is a critical milestone for detecting *Cirlivirales* in various datasets and exploring their potential role in AD and other neurodegenerative diseases.

4. Skills and Knowledge Gained

Throughout this internship, I have developed several key skills, including:

- **Bioinformatics Analysis:** I gained proficiency in tools such as AlphaFold3 for protein structure prediction and BLAST for sequence querying.
- **Database Management:** I learned how to construct and curate a specialized sequence database, focusing on reducing redundancy and ensuring comprehensive coverage.
- **Structural Biology:** I deepened my understanding of viral protein structures, particularly the jellyroll fold, and its significance in virus identification.

On a personal level, the biggest thing I learned is that I definitely want to continue working in bioinformatics. I've always been fascinated by how computer science and computational biology interact, and this experience has only deepened my interest. So, the short answer is yes—I'm excited to keep exploring this field and its connections.

5. Challenges and Solutions

I encountered many challenges, but I would say the main challenge was refining the database to ensure comprehensive coverage and accuracy. Some capsid protein sequences didn't align due to diversity or sequence length, which I addressed by adjusting the MSA process.

6. Conclusion

This internship provided valuable experience in bioinformatics and structural biology. The creation of the *Cirlivirales* jellyroll database is a significant achievement, laying the groundwork for future research into the role of these viruses in neurodegenerative diseases like AD. I am eager to continue exploring the connections between viral infections and human diseases.

7. Acknowledgments

I would like to thank King Abdullah University of Science and Technology for sponsoring me and giving me the opportunity to train at the prestigious University of Toronto. I also want to express my gratitude to Artem Babaian from the RNA lab for guiding me throughout the internship, ensuring I learned and solved problems on my own rather than providing immediate answers.

8. References

1. Seaks, C. E., & Wilcock, D. M. (2020). Infectious hypothesis of Alzheimer disease. *PLoS Pathogens*, 16(11), e1008596. https://doi.org/10.1371/journal.ppat.1008596
2. Eimer, W. A., et al. (2018). Alzheimer's Disease-Associated β-Amyloid Is Rapidly Seeded by Herpesviridae to Protect against Brain Infection. *Neuron*, 99(1), 56–63.e3. https://doi.org/10.1016/j.neuron.2018.06.030
3. Edgar, R. C. (2021). MUSCLE v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping. *bioRxiv* 2021.06.20.449169. https://doi.org/10.1101/2021.06.20.449169
4. Abramson, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630, 493–500. https://doi.org/10.1038/s41586-024-07487-w
5. HMMER: biosequence analysis using profile hidden Markov models. hmmer.org.
6. Nomburg, J., et al. (2024). Birth of new protein folds and functions in the virome. *bioRxiv* 2024.01.22.576744. https://doi.org/10.1101/2024.01.22.576744