Chapter 1 INTRODUCTION

1.1 Background

SARS-CoV-2, a novel coronavirus, was first discovered in December 2019 and quickly spread throughout worldwide (Wu et al., 2020). The global impact of SARS-CoV-2 is evident, 761,071,826 reported cases and 6,879,677 deaths as of 20 March 2023(WHO, 2023). The emergence of highly transmissible variants, including Omicron and Delta, has further higlighted the pressing need to comprehend the sequence diversity of SARS-CoV-2. These variants harbor specific mutations that may affect the virus's ability to evade immune responses or respond to treatments.

The remarkable mutation rate and consequent sequence diversity of SARS-CoV-2 have posed significant challenges the development of effective vaccines, therapeutics, and diagnostics. The immense sequence diversity obseved among viral populations hampers the design of robust effective surveillance and intervention strategies. The study of sequence diversity is a valuable approach for characterizing the evolutionary dynamics of viruses and for devising targeted interventions. However, the analysis of large sequence dataset poses a significant challenge to alignment-dependent approaches, which are typically used for sequence-based comparison studies. Multiple sequence alignment becomes impractical when expanding the encompass to include all species under a higher rank of the taxonomic lineage. As an alternative, alignment-free approaches have garnered increasing attention due to their efficiency in analyzing large sequence datasets.

Recently, we developed UNIQmin, a tool that employs an alignment-free approach for the study of viral sequence diversity at any given rank of taxonomy lineage. UNIQmin performs an exhaustive search to generate a minimal set for a given sequence dataset of interest, resulting in the smallest possible number of distinct sequences required to represent a given peptidome diversity exhibited by the dataset. This compression is achieved through the removal of sequences that do not contribute effectively to the peptidome diversity pool. UNIQmin has been demonstrated to be effective for several viral datasets, including species Dengue virus, genus Flavivirus, family Flaviviridae, and the superkingdom Viruses. In this study, we applied UNIQmin to protein sequence data of SARS-CoV-2 and its variants, to evaluate the effective viral sequence diversity at each rank. The utilization of UNIQmin on COVID-19 data has the potential to offer significant contributions in various aspects of research and analysis related to the SARS-CoV-2 virus. The program's ability to reduce protein sequences while preserving essential biological information can lead to valuable insights and advancements in the understanding and management of the SARS-CoV-2.

1.2 Project scope

Viruses, including SARS-CoV-2, are characterised by their high mutation rates, which surpass other pathogens (REF). This unique characteristics gives rise to dynamic ‘clouds’ of mutants, leading to reduced sequence conservation and ultimately resulting in a high degree of antigenic diversity. The extensive antigenic diversity exhibited by SARS-CoV-2 holds significant implications for vaccine efficacy, potentially impacting the effectiveness of immunization strategies. In this study, the UNIQmin is explored within the context of the spike protein of SARS-CoV-2 to address the following intriguing questions:

1. How can the computational performance of UNIQmin be enhanced through the implementation of parallel computing technique?
2. What are the differences in peptidome diversity within the spike protein of SARS-CoV-2 between the timeframes of July 2021 and December 2022?
3. How does the peptidome sequence diversity of the spike protein vary across different SARS-CoV-2 variants, specifically Alpha, Beta, Delta, Gamma, Mu, and Omicron, when analyzed using UNIQmin?

The fundamental concept underlying the minimal set approach is to extract the viral peptidome, representing the complete antigenic repertoire, from all reported sequences, and to determine the smallest number of sequences required to effectively represent the peptidome. Decoding the minimal set of the viral peptidome will provide valuable insights into the structure, function and evolution of SARS-CoV-2, with direct applications in diagnostics, therapeutics, and discovery of vaccine targets at the subgroup, species, and/or genus level(s). There is limited literature that is directly relevant to the study proposed herein. This alignment-free approach has been demonstrated to be effective for several viral datasets, including species Dengue virus, genus Flavivirus, family Flaviviridae, and the superkingdom Viruses (REF). Based on the aforementioned questions, we hypothesise herein that the spike protein of SARS-CoV-2 comprises of a limited number of core peptides that serve as the backbone for the combinatorial diversity of the millions of publicly reported sequences. The specific aims of this study are as follows:

1. To improve UNIQmin in terms of computational performance by implementing parallel computing.
2. To compare the peptidome diversity of spike protein between the retrieved dataset of July 2021 and December 2022.
3. To employ UNIQmin for the explore peptidome sequence diversity within the spike protein across different SARS-CoV-2 variants, including Alpha, Beta, Delta, Gamma, Mu, and Omicron, elucidating the variations in antigenic profiles.

By achieving these aims, this study seeks to advance our understanding of viral diversity within the SARS-CoV-2 spike protein and its variants. The insights gained from this research will contribute to the development of more effective strategies for diagnostics, therapeutics, and the identification of vaccine targets, ultimately aiding in the global response to the COVID-19 pandemic.