Scholar's Name: A S Md Mukarram Hossain

Thesis Title: Scalable Tools for High-throughput Viral Sequence Analysis

Email: mukarram819@yahoo.com

Home institute address: Department of Computer Science and Engineering, University of Dhaka, Dhaka – 1000, Bangladesh.

Host institute address: Department of Veterinary Medicine, University of Cambridge, Cambridge, CB3 OES, UK.

Abstract:

Viral sequence data are increasingly being used to estimate evolutionary and epidemiological parameters to understand the dynamics of viral diseases. This thesis focuses on developing novel and improved computational methods for high-throughput analysis of large viral sequence datasets. I have developed a novel computational pipeline, Pipelign, to detect potentially unrelated sequences from groups of viral sequences during sequence alignment. Pipelign detected a large number of unrelated and mis-annotated sequences from several viral sequence datasets collected from GenBank. I subsequently developed ANVIL, a machine learning-based recombination detection and subtyping framework for pathogen sequences. ANVIL's performance was benchmarked using two large HIV datasets collected from the Los Alamos HIV Sequence Database and the UK HIV Drug Resistance Database, as well as on simulated data. Finally, I present a computational pipeline named Phlow, for rapid phylodynamic inference of heterochronous pathogen sequence data. Phlow is implemented with specialised and published analysis tools to infer important phylodynamic parameters from large datasets. Phlow was run with three empirical viral datasets and their outputs were compared with published results. These results show that Phlow is suitable for high-throughput exploratory phylodynamic analysis of large viral datasets. When combined, these three novel computational tools offer a comprehensive system for large scale viral sequence analysis addressing three important aspects: 1) establishing accurate evolutionary history, 2) recombination detection and subtyping, and 3) inferring phylodynamic history from heterochronous sequence datasets.