Extended Abstract - Towards a phylogenetic measure to quantify HIV incidence

Pieter Libin *[1,2], Nassim Versbraegen*[5,6], Ana B. Abecasis[2,3], Perpetua Gomes[4], Tom Lenaerts[1,5], and Ann Nowé[1]

[1]Artificial Intelligence Lab, Department of computer science, Vrije Universiteit Brussel, Brussels, Belgium
[2]Department of Microbiology and Immunology, Rega Institute for Medical Research, KU Leuven - University of Leuven, Leuven, Belgium
[3]Global Health and Tropical Medicine, GHTM, Instituto de Higiene e Medicina Tropical, IHMT, Universidade Nova de Lisboa, UNL, Lisboa, Portugal
[4]Laboratorio Biologia Molecular, LMCBM, SPC, HEM, Centro Hospitalar Lisboa Ocidental
[5]Machine Learning group, Université Libre de Bruxelles, Boulevard du Triomphe CP212, 1050 Bruxelles, Belgium
[6]Interuniversity Institute for Bioinformatics in Brussels, ULB-VUB, 1050 Brussels, Belgium

About 37 million people are currently infected with HIV and an estimated 35 million people have died due to the effects of AIDS (the eventual result of an untreated HIV infection) since the beginning of the epidemic at the start of the twentieth century. Global efforts have ensued to enhance the collection, dissemination and accessibility of epidemiological data related to HIV epidemics. One of the most burdensome aspects in curtailing the spread of HIV emerges from infected individuals being unaware of their infection status. This stems from the fact that a human host can be infected for many years before noticing any symptoms. As a result, a significant fraction of the HIV infected population remains undiagnosed, hampering effectiveness of interventions and assessment of further developments of the epidemic. Consequently, methods that deliver a well-founded estimate of the number of HIV infected individuals are paramount.

[*] Equal contribution.

Such an estimate enables deduction of the number of undiagnosed infected individuals. State-of-the-art methods that aim to provide estimates of the size of HIV epidemics generally consist of applying compartment models to routine surveillance data to estimate the number of infected individuals (i.a., number of new diagnoses over time and CD4$^+$ cell counts).

An abundance of clinical data is available in the context of HIV epidemics, as upon diagnosis a number of tests are performed and the results thereof collected. One of those tests comprises the assessment of the genotype of the virus that infects the patient. To that extent, the genetic sequence of the virus is determined. As a result, a vast amount of HIV sequences have been collected over the last decades.

The main benefit of developing a method to quantify a HIV epidemic that relies on genetic data is to gain insight into the specific sub-populations that contain a high rate of undiagnosed individuals, thus allowing for more effective health policies, through diagnosis strategies that are directed towards these particular sub-populations.

We validate our research on the HIV-1 epidemic in Portugal. We therefore first present inference of the epidemiological parameters of said epidemic by applying approximate Bayesian computation (ABC). We apply ABC to fit a model that incorporates the epidemiological parameters in question. We further show that calibrating simulations to specific epidemics is essential, as the epidemiological dynamic has an important impact on the shape of the phylogenetic tree. We then construct a tree statistic that enables differentiation of epidemiological parameters based on phylogenetic trees and evaluate it on a set of epidemiological simulations. We show that the presented tree statistic enables differentiation of epidemiological parameters, while only relying on phylogenetic trees, thus enabling the construction of new methods to ascertain the epidemiological state of an HIV epidemic. By using genetic data to infer epidemic sizes, we expect to enhance understanding of the portions of the infected population in which diagnosis rates are low.