INFECTIOUS DISEASES

Effects of mutations on SARS-CoV-2 fitness

Science **376**, 1327–1332 (2022)

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic has been characterized by the emergence of new variants with increased fitness. Identifying new lineages and estimating their relative fitness is therefore an important task for guiding outbreak response. However, current phylogenetic approaches do not scale well when considering large numbers of genomes; for instance, more than 6 million genomes are available for SARS-CoV-2. In addition, estimates of relative fitness do not usually take into account the effects of individual mutations: this could help identify which mutations are more relevant to fitness increase, allowing for a better prediction and understanding of the biology of transmission. In order to address these gaps, Fritz Obermeyer, Jacob E. Lemieux and colleagues propose a hierarchical regression model that can infer lineage fitness, predict the fitness of new lineages, forecast future lineage proportions, and estimate the effects of individual mutations on fitness.

To avoid a full phylogenetic inference, the proposed model — PyR_0 — first clusters genomes by genetic similarity. After clustering, sequence data are used to construct spatiotemporal lineage prevalence counts and amino acid mutation covariates. Then, the authors fit these data to a Bayesian multivariate logistic multinomial regression model: the model regresses lineage counts against mutation covariates. The authors leverage PyTorch, a machine learning framework, and Pyro, a probabilistic programming language, to scale efficiently to large datasets.

The authors applied PyR_0 to all publicly available SARS-CoV-2 genomes, estimating the contribution of various mutations to lineage fitness. Driver mutations in the Spike protein, previously established experimentally, were identified by the model, and other non-Spike mutations were associated with the elevated fitness of BA.2 (an Omicron subvariant). In addition, the authors demonstrated the model's capability in inferring fitness of new mutations based on the trajectories of other lineages in which they have previously emerged. According to the authors, inference and prediction takes about ten minutes on a single graphics processing unit. The overall scalability and predictive features of the proposed model have the potential to better aid public health efforts in rapidly detecting and understanding new lineages as they emerge.

Fernando Chirigati

Published online: 19 July 2022
https://doi.org/10.1038/s43588-022-00289-y