Lay Abstract

We plan to distill evolutionary constraints from rapidly expanding databases (GISAID & NCBI) of $> 10,000$ SARS-CoV-2 sequences, to predict epitopes and sequences of perturbed fusion proteins expected to emerge in future in the wild. Our central idea in this project is to model the constraints on the variations of the nucleotide sequences as a virus evolves by inferring a set of inter-dependent predictors known as the Quasinet or the Enet. The Enet framework is specifically designed for the analysis of biological sequences at scale, with the objective of modeling and prediction of dynamics unfolding in ultra-high dimensional sequence spaces. The key idea here is surprisingly simple: *we learn models for predicting the mutational variations at each index of the genomic sequence using other indices as features. Collectively, these predictors represent the emergent constraints that shape evolutionary changes from selection forces in the wild.*