Deep learning for bacterial population genetics

Laboratoire de Recherche en Informatique, U. Paris Sud, CNRS

Surveillance of bacterial populations is more and more important given the rise of bacterial pathogen carrying antibiotic resistance genes. This surveillance can guide public health intervention, by monitoring the effect of a new treatment [1]. A goal of population genetics is to infer past demography of a population along with detecting the effect of natural selection on it. Summary statistics were developed to describe the raw genetic data into meaningful metrics to understand a population’s past history. However, these expert statistics do not grasp the totality of the information available, and thus have limited inference power, especially in complex scenario. Different methods for population history inference already exist. It has been shown that particularly in the case of bacterial population, they cannot infer the correct demographic history, because different effects such as selection, recombination or sampling blur the signal of demography [2]. To leverage the full potential of the data while taking into account the complexity of the signal, we develop a method inferring jointly demography and selection with a deep learning approach based on raw genetic data.

Population genetics datasets are rare and when they exist, the ground truth of the history of the population is unknown. Thus, we rely on simulated datasets to train a supervised method. Here we use the forward simulator Slim2 [3] with our implementation of bacterial populations which were not available before. We simulate a single population enduring a bottleneck and the effect of a mutation under positive selection.

Forward simulations are computationally expensive, thus we scale down the size of our target population by a given factor, while increasing the mutation and recombination rates by the same factor. We perform different simulations with different scaling factor to make sure the it doesn’t affect various summary statistics.

The parameters of the simulation are based on the bacterial population of a virulent clone of *Streptococcus agalactiae*.

---

**Bacterial population genetics**

---

**Deep learning for bacterial population genetics**

---

**Deep learning approach**

---