Deep learning for bacterial population genetics

J. Cury, T. Sanchez, G. Charpiat, G. Achaz, P. Glaser, E. Rocha, F. Jay
Laboratoire de Recherche en Informatique, U. Paris Sud, CNRS

Introduction

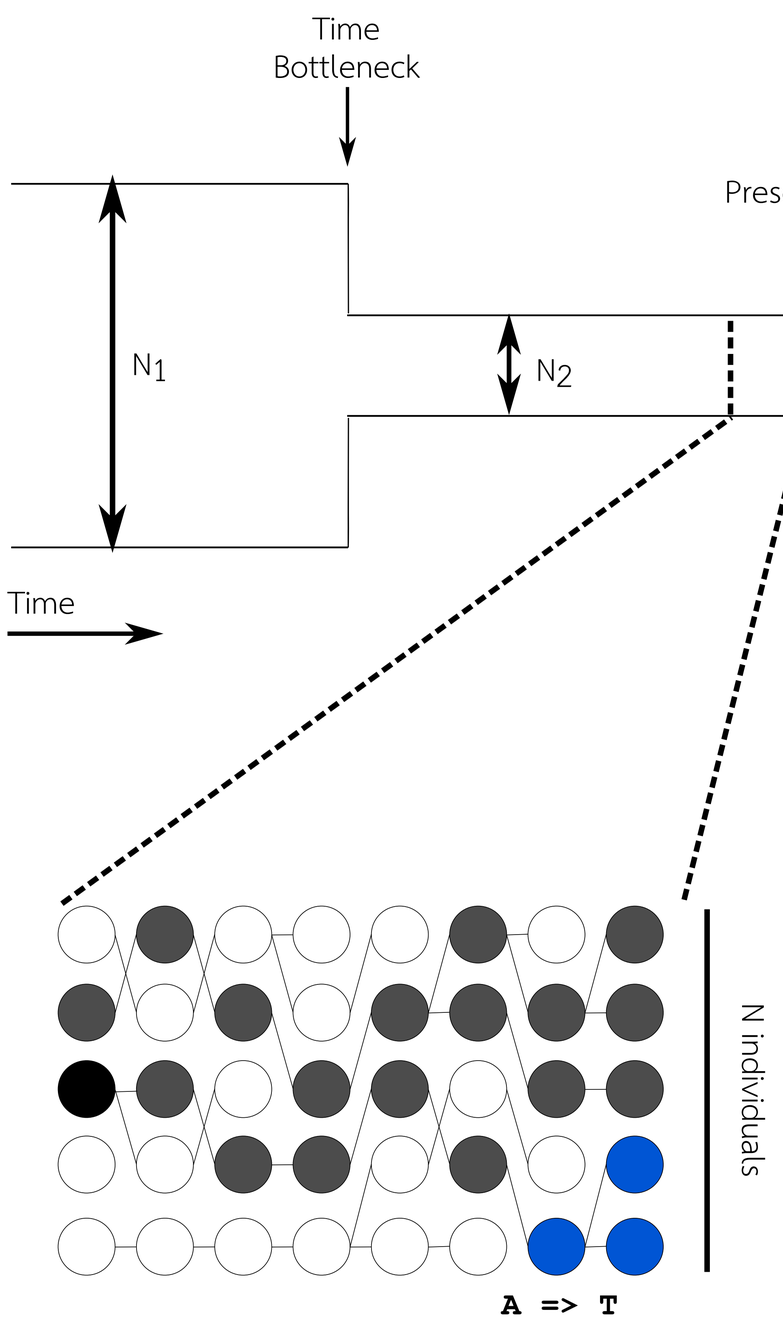
Surveillance of bacterial populations is more and more important given the rise of bacterial pathogen carrying antibiotic resistance genes. This surveillance can guide public health intervention, by monitoring the effect of a new treatment [1]. A goal of population genetics is to **infer past demography** 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.

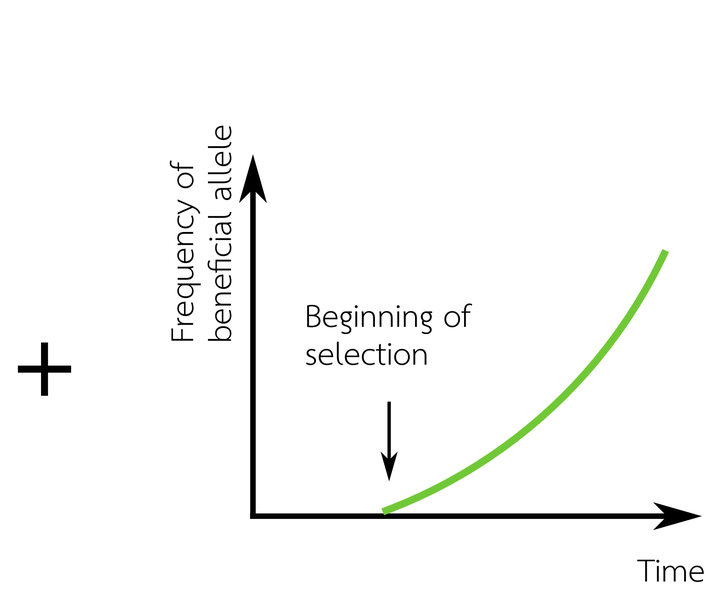
Bacterial population genetics

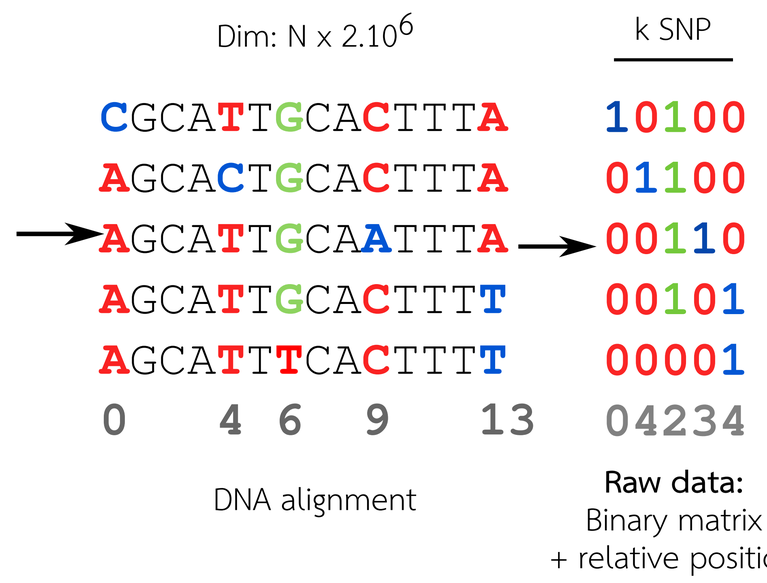
Population genetics datasets are rare and when they exist, the ground truth of the history of the population is **unknown**. Thus, we rely on **simulated datasets** to train a supervised method. Here we use the forward simulator Slim2 [3] with our implementation of bacterial populations which were not 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***.



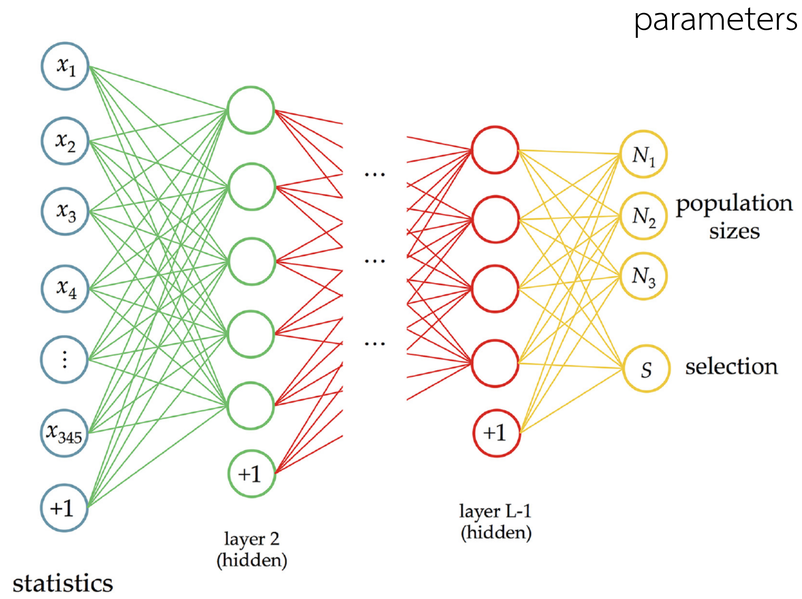




Goal: infer the **size of the population** (N_1 and N2 above), the **time** of demographic events, such as the bottleneck, but also to detect the **intensity** of the selective force and the **position** of the locus under selection.
Such task is particularly complex because a bottleneck or a selection event can lead to the **same signals**, such as the increase of mutations present in the majority of individuals. Besides, the imprint of demography is **globally** spread along the genome while the selection signal is **locally** detectable around the locus under selection.
A deep learning approach has already been developed to solve this task, but the input was a few hundreds of summary statistics (MLP on the left) [4].
Here we are developping an **deep convolutional architecture** which takes as input the raw data. The raw data ressembles to an image but is structured only on the dimension of the SNP, and not on the dimension of the individuals.

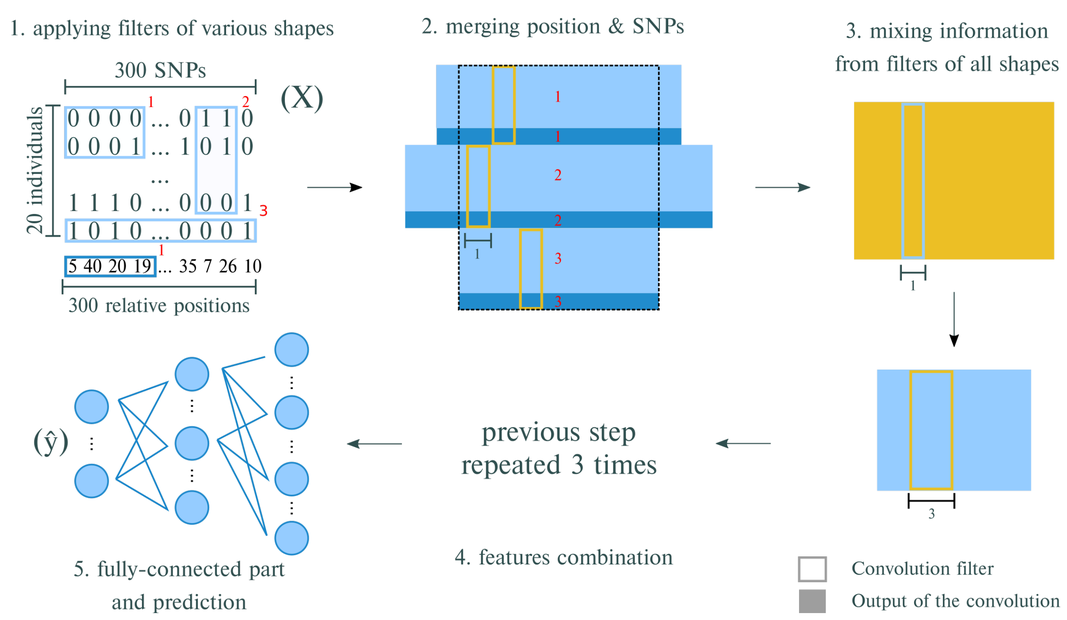
Deep learning

Summary statistics → Demography + Selection parameters



Multilayer Perceptron to infer population size and selection with summary statistics

Raw data → Demography parameters



Convolutional Artificial Neural Network to infer demographic scenario

Ref.

[1] N. J. Croucher et al., "Evidence for Soft Selective Sweeps in the Evolution of Pneumococcal Multidrug Resistance and Vaccine Escape," Genome Biol Evol, 2014.
[2] M. Lapierre, C. Blin, A. Lambert, G. Achaz, and E. P. C. Rocha, "The Impact of Selection, Gene Conversion, and Biased Sampling on the Assessment of Microbial Demography," Mol. Biol. Evol., 2016
[3] B. C. Haller and P. W. Messer, "SLiM 2: Flexible, Interactive Forward Genetic Simulations," Mol. Biol. Evol., 2017.
[4] S. Sheehan and Y. S. Song, "Deep Learning for Population Genetic Inference," PLOS Comput. Biol. 2016.