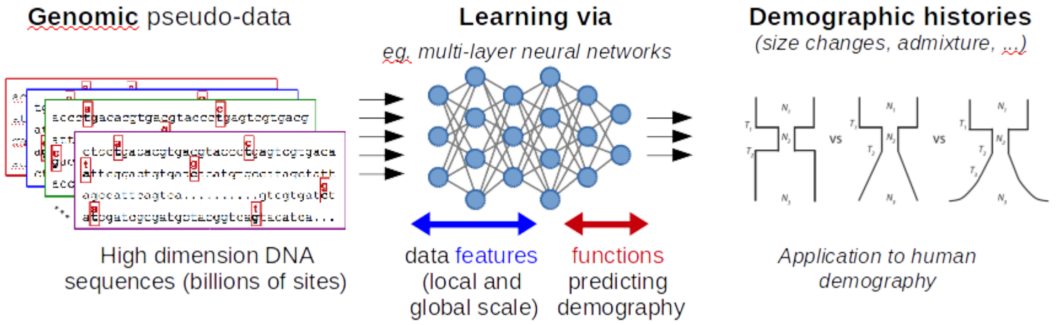
Reconstructing the past: deep learning for population genetics

Flora Jay (Bioinfo, LRI) and Guillaume Charpiat (TAO, LRI)
flora.jay@lri.fr guillaume.charpiat@inria.fr

Context: With the never-ending improvement of sequencing technology, more and more genetic data is available, covering either millions of DNA sites or thousands of individuals, and sometimes both. This massive data should greatly enhance our knowledge about **past evolution and population demographic history.** This history can indeed be reconstructed over the past thousands of years thanks to the inference of present-day individual relationships, by comparing their DNA, identifying shared genetic mutations, their frequency, and their correlation at different genomic scales. In turn this leads to a better understanding of populations, such as their fluctuation in size over time, the timing of divergence and migration between two groups, or the adaptation to their local environment. However the best way to extract information from large genomic data is still an open problem and for now is mostly achieved through a drastic reduction of dimension to a few well-studied population genetics features.

Goals: We aim to develop a new method for demographic inference from genomic data, ie for estimating relevant parameters, such as the population sizes (number of individuals) at different epochs. Our approach will have common roots with Approximate Bayesian Computation which consists in simulating numerous pseudo-datasets according to specified models (here demographic) with prior distributions, and comparing predefined population genetics statistics in pseudo and real data. In our case, we aim at **learning the complex relationship linking genomic data to demographic models** using machine learning techniques rather than relying on predefined statistics. First, we will evaluate the feasibility of applying deep learning methods that were proven useful for image and text analysis, such as convolution and recurrent networks, and **build a deep learning architecture tailored to our genomic data and its specific invariances.** The most notable invariances are the invariance by permutation over individuals and by translation over the DNA sequence, but other data characteristics exist and will be described mathematically. A second challenge will be to **develop a flexible architecture that can handle input data with different sizes.** Contrary to images of fixed dimension, genetic data might vary both in term of number of sequenced individuals and number of sites. We suggest two approaches: one consists in building and training a meta neural network that could generate architectures tailored for each input size, while the other focuses on picking families of functions naturally extendable to any input size. In the second approach, neurons described by these functions could be spread in the network enabling automatic adaptation to the size of the previous layer. Finally, we will apply the method to human genetics data and compare our findings to current knowledge in anthropology genetics.


Genomic pseudo-data
High dimension DNA sequences (billions of sites)
Learning via
eg. multi-layer neural networks
data features (local and global scale)
functions predicting demography
Demographic histories
(size changes, admixture, ...)
vs
vs
Application to human demography

Bibliography

. Boitard, et al. Inferring population size history from large samples of genome-wide molecular data - an approximate Bayesian computation approach. *PLoS Genet*. (2016)

. Sheehan S, Song YS. Deep learning for population genetic inference. *PLoS Comput Biol*. (2016)

. Stanley, et al. A hypercube-based encoding for evolving large-scale neural networks. *Artificial life* 15.2 (2009)