Deep learning algorithms in evolutionary genomics: detecting signatures of natural selection

Lorenzon L(1,2)† , Pattini L(1), Mathieson S(3), Fumagalli M(2)

(1) Department of Life Sciences, Imperial College London, London, UK
(2) Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy
(3) Department of Computer Science, Swarthmore College, Swarthmore, PA, USA

† Email: m.fumagalli@imperial.ac.uk

Motivation
The genetic basis of complex phenotypes, including susceptibility to certain diseases such as Type 2 Diabetes, are still largely unknown due to the polygenic nature of the trait and the small effect each associated mutation contributes to it. Conversely to classic association studies, adopting an evolutionary approach is a promising strategy to unveil functional mutations in the human genome. In fact, as sites targeted by natural selection are likely to harbor important functionalities for the carrier, the identification of selection signatures in the genome has the potential to elucidate the genetic mechanisms underpinning human phenotypes, including susceptibility to complex disorders. Commonly used strategies to detect such signals rely on compressing the genomic information into a set of summary statistics, whose expected distributions are assessed analytically or empirically. While popular, these methods inevitably rely on a significant loss of information and their power is limited to simple scenarios of selective regimes. Here we explore the use of deep learning to make full use of population genomics data into an evolutionary-epidemiological context.
Methods
Our approach is based on translating the information of genome variation from multiples individuals and populations into abstract images. Specifically, each population image is created by stacking aligned genetic data and encoding derived and ancestral alleles as black and white pixels, respectively. Rows are then ordered by increasing frequency of occurrence to highlight patterns of randomness under neutral evolution (Figure 1A) or decreased variability under natural positive selection (Figure 1B). We implemented a Convolutional Neural Network to recognise patterns of natural selection from random genetic drift, and classify genes into selectively or neutrally evolving. We trained the network using extensive simulations under various evolutionary regimes to optimise the weights connecting nodes and layers in the network. We finally calculated the prediction accuracy of prediction from a validation set of images.
Results
In the case of binary classification of a genomic region experiencing neutral evolution or weak-to-moderate positive selection, our method perfectly distinguishes between them with a prediction accuracy over 99%. Also, the method seems to be

robust to deviations from the underlying demographic model. We then attempted to perform a classification into 41 classes representing the strength of the signal, in the form of the selection coefficient. Under this framework, the final weight scores are an approximation for the posterior distribution of our parameter of interest. Our network is able to accurately estimate the selection coefficient with an average root mean square deviation of 0.0016. (Figure 1C). We finally demonstrate the utility of this tool to detect positive selection in a set of candidate genes associated to cardiovascular disorders, including hypertension.

While the use of deep learning in evolutionary genomics is at its infancy, here we demonstrate its potentiality to detect informative patterns from large-scale genomic data. The joint use of inferences of the evolutionary history of mutations and their functional impact will facilitate mapping studies and provide novel insights on the molecular mechanisms leading to disorders.