Identification of Natural Selection in Genomic Data with Deep Convolutional Neural Network

Arnaud Nguembang Fadjia (✉ arnaud.nguembafadja@unife.it)

University of Ferrara: Università degli Studi di Ferrara <https://orcid.org/0000-0003-3352-5260>

Fabrizio Riguzzi

University of Ferrara Department of Mathematics: Università degli Studi di Ferrara Dipartimento di Matematica e Informatica

Giorgio Bertorelle

Università degli Studi di Ferrara Dipartimento di Scienze della Vita e Biotecnologie

Emiliano Trucchi

Università Politecnica delle Marche Dipartimento di Scienze della Vita e dell'Ambiente: Università Politecnica delle Marche Dipartimento di Scienze della Vita e dell'Ambiente

Research

Keywords: Genomic data, Inference of natural selection, Deep Learning, Convolutional Neural Networks

Posted Date: June 22nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-624734/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: With the increase in the size of genomic datasets describing variability in populations, extracting relevant information becomes increasingly useful as well as complex. Recently, computational methodologies such as Supervised Machine Learning and specifically Convolutional Neural Networks have been proposed to order to make inferences on demographic and adaptive processes using genomic data, Even though it was already shown to be powerful and efficient in different fields of investigation, Supervised Machine Learning has still to be explored as to unfold its enormous potential in evolutionary genomics.

Results: The paper proposes a method based on Supervised Machine Learning for classifying genomic data, represented as windows of genomic sequences from a sample of individuals belonging to the same population. A Convolutional Neural Network is used to test whether a genomic window shows the signature of natural selection. Experiments performed on simulated data show that the proposed model can accurately predict neutral and selection processes on genomic data with more than 99% accuracy.

Full Text

This preprint is available for [download as a PDF](#).