Prediction of mutational effects in plants using artificial intelligence

Eva Pardo-Otero[1], Verónica Bolón-Canedo[2], and Rosalía Piñeiro[3]

[1] GIBE, CITIC, Universidade da Coruña, A Coruña, Spain
[2] LIDIA, CITIC, Universidade da Coruña, A Coruña, Spain
[3] GIBE, CICA, Universidade da Coruña, A Coruña, Spain

Abstract

Deleterious mutations can accumulate as genetic load in the genome, reducing biological fitness and compromising the viability of populations and species—an aspect still poorly explored in plants. Artificial intelligence (AI) offers new opportunities to predict mutational effects and refine these estimates. We evaluated the performance of classical and AI-based methods for predicting mutational effects using 12,865 variants from 6,172 protein sequences across 433 plant species and explored the use of embeddings derived from protein language models (PLMs) to train supervised models. These advances may contribute to a more accurate estimation of genetic load, a key factor for guiding conservation strategies and the sustainable management of natural resources.

1 Introduction

Genetic load refers to the accumulation of deleterious mutations in the genome that reduces fitness, an aspect still poorly explored in plants [1]. In small populations, reduced efficacy of purifying selection allows mildly deleterious mutations to accumulate, increasing the load and compromising the viability of populations and species. Estimating genetic load is therefore essential for designing effective conservation strategies, including genomic rescue and managed gene flow, aimed at maintaining long-term viability. To this end, accurately predicting mutational effects is key to identifying deleterious mutations and estimating the genetic load.

Classical approaches like SIFT [3] rely on sequence evolutionary conservation to infer functional importance. Protein language models (PLMs) like ESM1v [4] learn evolutionary and structural patterns from large protein databases. Once pretrained, they can predict the impact of mutations in a zero-shot setting by comparing the likelihood the model assigns to the mutant versus the wild-type residue. Alternatively, PLMs can extract high-dimensional embeddings that represent protein properties and can be used in supervised models trained for mutational effect prediction. In contrast to SIFT, which evaluates mutational effects from position-specific conservation scores, PLMs take into account the surrounding aminoacid context and capture how mutations can disrupt sequence dependencies. These AI-based methods remain largely unexplored in plants, where predictive tools are still limited.

In this work, we compare three approaches to predict mutational effects in plant proteins: a classical non-AI-method (SIFT), a zero-shot PLM (ESM1v), and a supervised model using ESM1v embeddings. Employing 12,865 variants from 6,172 proteins of 433 plant species [2], we evaluate predictive accuracy and environmental impact.

2 Methodology

Performance evaluation was conducted using the corresponding test set (20%). SIFT used default settings and Viridiplantae-filtered UniRef90 database. The PLM zero-shot approach used the average predictions from five ESM1v ensemble models. The PLM embedding-based supervised approach used 1,280-dimensional ESM1v embeddings of the mutated aminoacid residues as input features for a Support Vector Machine (SVM) trained on the corresponding training set (80%). All analyses ran on CESGA's FinisTerrae III HPC system, using 16 CPU cores for SIFT and an A-100 GPU for ESM1v. Energy use and emissions were estimated with CodeCarbon v3.0.7, based on hardware usage and regional carbon intensity.

3 Results

The comparative evaluation of the three approaches revealed distinct performance and environmental impact (Table 1). The embedding-based supervised approach achieved the highest f1-score of 0.819 and balanced performance. The classical SIFT method performed moderately well (f1-score = 0.754) and showed the highest ability to detect deleterious mutations (recall = 0.872), but at the cost of lower precision of 0.710. In contrast, the ESM1v zero-shot model prioritizes minimizing false positives, achieving a higher precision of 0.802, but missing true deleterious mutations (recall = 0.625). Regarding energy consumption, SIFT is the fastest and most efficient option, with 40× less energy and lower CO_2 emission. However, the ESM1v models require more computation and therefore produce higher emissions, mainly due to the embedding extraction phase in the case of the supervised approach.

Methods	**F1-Score**	**Precision**	**Recall**	**Time (s)**	**E (Wh)**	**CO_2 (g)**
SIFT	0.757	0.710	**0.872**	**76**	**6.21**	**1.08**
ESM1v zero-shot	0.734	0.802	0.625	2358	255.47	44.47
ESM1v supervised	**0.819**	**0.806**	0.835	2013	197.23	34.32
**embedding extraction*				1983	193.85	33.74
training				26	3.01	0.52
testing				4	0.37	0.06

Table 1: Comparison of performance. Runtime, energy use and CO_2 emissions inferred from processing a subset of 100 mutations. Best values in **bold**. *Embedding extraction of both train and test samples.

References

[1] Giorgio Bertorelle, Francesca Raffini, Mirte Bosse, Chiara Bortoluzzi, Alessio Iannucci, Emiliano Trucchi, Hernán E. Morales, and Cock van Oosterhout. Genetic load: genomic estimates and applications in non-model animals. *Nature Reviews Genetics*, 23:492–503, 8 2022.

[2] Xiangjian Gou, Xuanjun Feng, Haoran Shi, Tingting Guo, Rongqian Xie, Yaxi Liu, Qi Wang, Hongxiang Li, Banglie Yang, Lixue Chen, and Yanli Lu. Ppved: A machine learning tool for predicting the effect of single amino acid substitution on protein function in plants. *Plant Biotechnology Journal*, 20:1417–1431, 7 2022.

[3] Prateek Kumar, Steven Henikoff, and Pauline C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature Protocols*, 4:1073–1082, 2009.

[4] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 12 2021.