

---

# Machine Learning para corrección de errores en datos de secuenciación de ADN

Trabajo de Fin de Grado

Amin Kasrou Aouam



**UNIVERSIDAD  
DE GRANADA**

26-06-2021

## Índice

|  |          |
|--|----------|
| <b>Resumen</b>                           | <b>3</b> |
| <b>Abstract</b>                          | <b>3</b> |
| <b>Introducción</b>                      | <b>4</b> |
| <b>Motivación</b>                        | <b>4</b> |
| <b>Estado del arte</b>                   | <b>4</b> |
| NGS . . . . .                            | 4        |
| Bioinformática (Deep Learning) . . . . . | 4        |
| <b>Metodología</b>                       | <b>4</b> |
| Tecnologías . . . . .                    | 4        |
| Pipeline . . . . .                       | 4        |
| Reproducibilidad . . . . .               | 4        |
| <b>Resultados</b>                        | <b>4</b> |
| <b>Conclusiones</b>                      | <b>4</b> |
| <b>Futuras mejoras</b>                   | <b>4</b> |

## Resumen

Las nuevas técnicas de secuenciación de ADN (NGS) han revolucionado la investigación en genómica. Estas tecnologías se basan en la secuenciación de millones de fragmentos de ADN en paralelo, cuya reconstrucción se basa en técnicas de bioinformática. Aunque estas técnicas se apliquen de forma habitual, presentan tasas de error significantes que son perjudiciales para el análisis de regiones con alto grado de polimorfismo. En este estudio se implementa un nuevo método computacional, locimend, basado en *Deep Learning* para la corrección de errores de secuenciación de ADN. Se aplica al análisis de la región determinante de complementariedad 3 (CDR3) del receptor de linfocitos T (TCR), generada *in silico* y posteriormente sometida a un simulador de secuenciación con el fin de producir errores de secuenciación. Empleando estos datos, entrenamos una red neuronal convolucional (CNN) con el objetivo de generar un modelo computacional que permita la detección y corrección de los errores de secuenciación.

## Abstract

Next generation sequencing (NGS) have revolutionised genomic research. These technologies perform sequencing of millions of fragments of DNA in parallel, which are pieced together using bioinformatics analyses. Although these techniques are commonly applied, they have non-negligible error rates that are detrimental to the analysis of regions with a high degree of polymorphism. In this study we propose a novel computational method, locimend, based on a *Deep Learning* algorithm for DNA sequencing error correction. It is applied to the analysis of the complementarity determining region 3 (CDR3) of the T-cell receptor (TCR), generated *in silico* and subsequently subjected to a sequencing simulator in order to produce sequencing errors. Using these data, we trained a convolutional neural network (CNN) with the aim of generating a computational model that allows the detection and correction of sequencing errors.

## **Introducción**

## **Motivación**

## **Estado del arte**

## **NGS**

## **Bioinformática (Deep Learning)**

## **Metodología**

## **Tecnologías**

## **Pipeline**

## **Reproducibilidad**

## **Resultados**

## **Conclusiones**

## **Futuras mejoras**