UNIVERSIDAD DE GRANADA

TRABAJO FIN DE GRADO

GRADO DE INGENIERÍA EN INFORMÁTICA

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


**Autor**

Amin Kasrou Aouam

**Directores**

Carlos Cano Gutiérrez

María Soledad Benítez Cantos




Facultad de Educación, Tecnología y Economía de Ceuta


—

Ceuta, 26 de Junio de 2021

# Índice general

# 1 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 detrimentales 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.

**Palabras clave**: deep learning, corrección de errores, receptor de linfocitos T, secuenciación de ADN, inmunología

# 2 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 polimorphism. 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.

**Keywords**: deep learning, error correction, DNA sequencing, T-cell receptor, immunology

# 3 Introducción

La secuenciación de ADN es el proceso mediante el cual se determina el orden de los nucleótidos en una secuencia de ADN. En los años 70, Sanger et al. desarrollaron métodos para secuenciar el ADN mediante técnicas de terminación de cadena. [1] Este avance revolucionó la biología, proporcionando las herramientas necesarias para descifrar genes, y posteriormente, genomas completos.

La demanda creciente de un mayor rendimiento llevó a una automatización y paralelización de las tareas de secuenciación.

La capacidad del sistema inmunitario adaptativo para responder a cualquiera de los numerosos antígenos extraños potenciales a los que puede estar expuesta una persona depende de los receptores altamente polimórficos expresados por las células B (inmunoglobulinas) y las células T (receptores de células T [TCR]). La especificidad de las células T viene determinada principalmente por la secuencia de aminoácidos codificada en los bucles de la tercera región determinante de la complementariedad (CDR3). [2]

## 3.1. Estado del arte

### 3.1.1. NGS

### 3.1.2. Bioinformática (Deep Learning)

# 4 Motivación y Objetivos

# 5 Metodología

## 5.1. Tecnologías

## 5.2. Pipeline

## 5.3. Reproducibilidad

# 6 Resultados

# 7 Conclusiones

# 8 Futuras mejoras

# Bibliografía

[1] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences*, vol. 74, no. 12, pp. 5463–5467, 1977, doi: 10.1073/pnas.74.12.5463.

[2] H. S. Robins *et al.*, "Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells," *Blood*, vol. 114, no. 19, pp. 4099–4107, Nov. 2009.