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 1

2 Abstract 2

3 Introducción 3

4 Estado del arte 5
- 4.1 Bioinformática 5
- 4.2 Deep Learning 5

5 Objetivos 6

6 Métodos 7
- 6.1 Tecnologías 7
- 6.2 Pipeline 7
- 6.3 Reproducibilidad 7

7 Resultados 8

8 Conclusiones 9

9 Futuras mejoras 10

Bibliografía 11

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 la automatización y paralelización de las tareas de secuenciación. Gracias a estos avances, la técnica de Sanger permitió determinar la primera secuencia del genoma humano en 2004 (Proyecto Genoma Humano). [2]

Sin embargo, el Proyecto Genoma Gumano requerió una gran cantidad de tiempo y recursos, y era evidente que se necesitaban tecnologías más rápidas, de mayor rendimiento y más baratas. Por esta razón, en el mismo año (2004) el *National Human Genome Research Institute* (NHGRI) puso en marcha un programa de financiación con el objetivo de reducir el coste de la secuenciación del genoma humano a 1000 dólares en diez años. [3] Esto estimuló el desarrollo y la comercialización de las tecnologías de secuenciación de alto rendimiento o *Next-Generation Sequencing* (NGS), en contraposición con el método automatizado de Sanger, que se considera una tecnología de primera generación.

Estos nuevos métodos de secuenciación proporcionan tres mejoras importantes: en primer lugar, en lugar de requerir la clonación bacteriana de los fragmentos de ADN, se basan en la preparación de bibliotecas de moléculas en un sistema sin células. En segundo lugar, en lugar de cientos, se producen en paralelo de miles a muchos millones de reacciones de secuenciación. Finalmente, estos resultados de secuenciación se detectan directamente sin necesidad de electroforesis. [4]

Actualmente, se encuentran en desarrollo las tecnologías de tercera generación de secuenciación (Third-Generation Sequencing). Existe un debate considerable sobre la diferencia entre la segunda y tercera generación de secuenciación, la secuenciación en tiempo real y la divergencia simple con respecto a las tecnologías anteriores deberían ser las características definitorias de la tercera generación. Aquí conside-

ramos que las tecnologías de tercera generación son aquellas capaces de secuenciar moléculas individuales, negando el requisito de amplificación del ADN que comparten todas las tecnologías anteriores. [5]

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). [6]

4 Estado del arte

4.1. Bioinformática

4.2. Deep Learning

5 Objetivos

6 Métodos

6.1. Tecnologías

6.2. Pipeline

6.3. Reproducibilidad

7 Resultados

8 Conclusiones

9 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] I. H. G. S. Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, pp. 931–945, Oct. 2004, doi: 10.1038/nature03001.

[3] J. A. Schloss, "How to get genomes at one ten-thousandth the cost," *Nature Biotechnology*, vol. 26, no. 10, pp. 1113–1115, Oct. 2008, doi: 10.1038/nbt1008-1113.

[4] E. L. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes, "Ten years of next-generation sequencing technology," *Trends in Genetics*, vol. 30, no. 9, pp. 418–426, Sep. 2014, doi: 10.1016/j.tig.2014.07.001.

[5] J. M. Heather and B. Chain, "The sequence of sequencers: The history of sequencing DNA," *Genomics*, vol. 107, no. 1, pp. 1–8, 2016, doi: https://doi.org/10.1016/j.ygeno.2015.11.003.

[6] 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.