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, Julio de 2021

Índice general

Índice de cuadros

Índice de figuras

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

El ácido desoxirribonucleico (ADN) y el ácido ribonucleico (ARN) son los repositorios moleculares de la información genética. La estructura de cada proteína, y en última instancia de cada biomolécula y componente celular, es producto de la información programada en la secuencia de nucleótidos de una célula. La capacidad de almacenar y transmitir la información genética de una generación a otra es una condición fundamental para la vida. Un segmento de una molécula de ADN que contiene la información necesaria para la síntesis de un producto biológico funcional, ya sea una proteína o un ARN, se denomina gen. El almacenamiento y la transmisión de información biológica son las únicas funciones conocidas del ADN. [1]

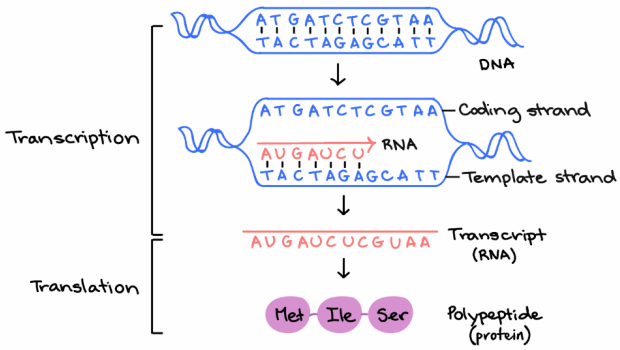

Figura 3.1: Dogma central de la biología molecular

Hay muy pocos principios firmes en biología. A menudo se dice, de una forma u otra, que la única regla real es que no hay reglas, es decir, que se pueden encontrar excepciones a cada principio fundamental si se busca lo suficiente. El principio conocido como el Dogma central de la biología molecular parece ser una excepción a esta regla de excepción ubicua. [2] El dogma central de la biología molecular establece que una vez que la información ha pasado a proteína no puede volver a salir; i.e. la transferencia de información de ácido nucleico a ácido nucleico, o de ácido nucleico a proteína puede ser posible, pero la transferencia de proteína a proteína, o de proteína a ácido nucleico es imposible. [3]

Por lo tanto, si elucidamos la información contenida en el ADN, obtenemos información sobre las biomoléculas que realizan las diferentes tareas fisiológicas y metabólica (e.g. ARN, proteínas).

3.1. Secuenciación de ADN

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. [4] 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). [5]

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. [6] 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.

3.2. Técnicas de secuenciación de alto rendimiento

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. [7]

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 consideramos 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. [8]

Estas nuevas técnicas han demostrado su valor, con avances que han permitido secuenciar el genoma humano completo, incluyendo las secuencias repetitivas (de telómero a telómero). Combinando los aspectos complementarios de las tecnologías Oxford Nanopore y PacBio HiFi, 2111 nuevos genes, de los cuales 140 son codificantes, fueron descubiertos en el genoma humano. [9]

3.3. Limitaciones de los métodos paralelos

Aunque las tecnologías de secuenciación paralelas (NGS) han revolucionado el estudio de la variedad genómica entre especies y organismos individuales, la mayoría tiene una capacidad limitada para detectar mutaciones con baja frecuencia. Este tipo de análisis es esencial para detectar mutaciones en oncogenes (genes responsables de la transformación de una célula normal a maligna), pero se ve restringido por una tasa de errores de secuenciación no despreciables. En 2011, la tasa de errores por substitución (intercambio de un nucleótido por otro) era > 0.1 %, y era similar en estudios posteriores. [10]

Para contrarrestar este obstáculo, varias técnicas mitigatorias se han puesto en marcha. Una de las más populares es el uso de una secuencia de consenso, que es un perfil estadístico a partir de un alineamiento múltiple de secuencias. Es una forma básica de descubrimiento de patrones, en la que un alineamiento múltiple de secuencias más amplio se resume en las características que se conservan. Este tipo de análisis permite determinar la probabilidad de cada base en cada posición de una secuencia. [11]

Todas las técnicas de consenso monocatenarias reducen los errores en dos o tres órdenes de magnitud, lo que es mucho mayor que cualquier enfoque computacional o bioquímico anterior, y permiten iden-

tificar con precisión variantes raras por debajo del 0.1 % de abundancia. Sin embargo, persisten algunos errores. Los errores que se producen durante la primera ronda de amplificación pueden propagarse a todas las demás copias escapando la corrección. [12]

Este problema se agrava en el análisis de repertorios inmunológicos, debido a nuestra limitada capacidad para distinguir entre la verdadera diversidad de los receptores de los linfocitos T (TCR) e inmunoglobulinas (IG) de los errores de PCR y secuenciación que son inherentes al análisis del repertorio. Los clonotipos resultantes pueden tener concentraciones drásticamente diferentes, lo que hace que los clonotipos menores sean indistinguibles de las variantes erróneas. [13]

3.4. Variedad genética en el sistema inmunitario

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). [14]

En el timo, durante el desarrollo de los linfocitos T, se selecciona al azar un segmento de cada familia mediante un proceso conocido como recombinación somática o recombinación V(D)J, de modo que se eliminan del genoma del linfocito los no seleccionados y los segmentos V(D)J escogidos quedan contiguos, determinando la secuencia de las subunidades del TCR y, por tanto, la especificidad de antígeno de la célula T. La selección aleatoria de segmentos junto con la introducción o pérdida de nucleótidos en sus uniones son los responsables directos de la variabilidad de TCR, cuya estimación es del orden de 10^{15} posibles especies distintas o clonotipos. [15]

Debido a la diversidad de uniones, las moléculas de anticuerpos y TCR muestran la mayor variabilidad, que forman la región determinante de la complementariedad 3 (CDR3). De hecho, debido a la diversidad de uniones, el número de secuencias de aminoácidos que están presentes en las regiones CDR3 de las de las moléculas de Ig y TCR es

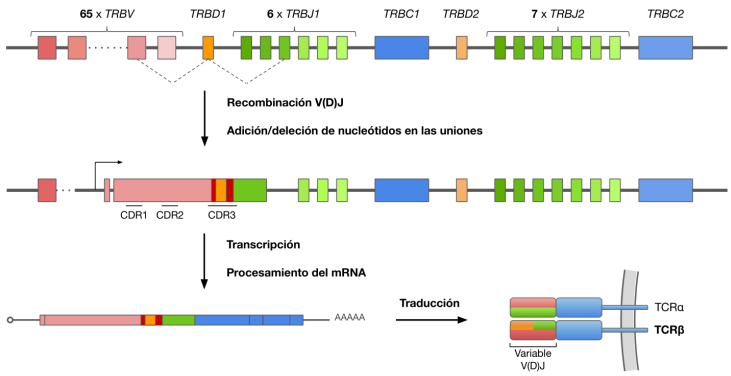

Figura 3.2: Generación de diversidad en el TCR α β. Durante el desarrollo de los linfocitos T se reordenan los segmentos génicos V (rosa), D (naranja) y J (verde) a través del proceso de recombinación V(D)J. Durante este proceso se pueden añadir o eliminar nucleótidos en la unión de los segmentos (rojo), contribuyendo a la diversidad de la secuencia. Se señalan las 3 regiones CDR, estando CDR3 localizada en la unión V(D)J. [15]

mucho mayor que el número que pueden ser codificadas por segmentos de genes de la línea germinal. [16]

Frente a la evidencia recaudada, diversos métodos computacionales basados en la inteligencia artificial se aplican para aliviar estos impedimentos.

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] M. M. C. Albert Lehninger David L. Nelson, *Lehninger-principles of biochemistry*, 5th Edition. W. H. Freeman, 2008, p. 276.

[2] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 1970, doi: 10.1038/227561a0.

[3] F. H. Crick, "On protein synthesis," in *Symp soc exp biol*, 1958, vol. 12, p. 8.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] S. Nurk *et al.*, "The complete sequence of a human genome," *bioRxiv*, 2021, doi: 10.1101/2021.05.26.445798.

[10] X. Ma *et al.*, "Analysis of error profiles in deep next-generation sequencing data," *Genome Biology*, vol. 20, no. 1, p. 50, Mar. 2019, doi: 10.1186/s13059-019-1659-6.

[11] C. Lee, "Generating consensus sequences from partial order multiple sequence alignment graphs," *Bioinformatics*, vol. 19, no. 8, pp. 999–1008, May 2003, doi: 10.1093/bioinformatics/btg109.

[12] J. J. Salk, M. W. Schmitt, and L. A. Loeb, "Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations," *Nature Reviews Genetics*, vol. 19, no. 5, pp. 269–285, May 2018, doi: 10.1038/nrg.2017.117.

[13] M. Shugay *et al.*, "Towards error-free profiling of immune repertoires," *Nature Methods*, vol. 11, no. 6, pp. 653–655, Jun. 2014, doi: 10.1038/nmeth.2960.

[14] 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.

[15] M. S. B. Cantos, "Análisis de repertorios de receptores de células t a partir de datos de secuenciación masiva," Master's thesis, Universidad de Granada, 2019.

[16] A. K. Abbas, A. H. Lichtman, and S. Pillai, in *Cellular and molecular immunology*, 9th ed., Elsevier, 2017, p. 204.