Analizando los avances y aplicaciones del Proyecto Genoma Humano

Ejercicio 2 – Bioinformática

Estudiante: Gisela Belmonte Cruz
Grado: Ciencia e Ingeniería de Datos – ULPGC
Curso: 2025

Artículo analizado:

Deorowicz, S., Danek, A. & Niemiec, M. (2015).
GDC 2: Compression of large collections of genomes.
Scientific Reports, 5:11565.

Disponible en Nature: https://www.nature.com/articles/srep11565
Versión en ResearchGate

Índice

1. Introducción

En los últimos años, el abaratamiento de las tecnologías de secuenciación genómica ha transformado el panorama de la biología y la medicina. Hoy es posible secuenciar genomas humanos completos de manera rápida y asequible, impulsando grandes proyectos internacionales como el *1000 Genomes Project* o el *UK10K*.

Sin embargo, este progreso científico genera un problema técnico importante: la cantidad de información producida por la secuenciación masiva es enorme y requiere soluciones de almacenamiento eficientes. Cada genoma humano ocupa varios gigabytes, por lo que conservar miles de ellos supone un reto para la infraestructura informática.

Para responder a este desafío, los investigadores Sebastian Deorowicz, Agnieszka Danek y Marcin Niemiec desarrollaron el algoritmo **GDC 2 (Genome Differential Compressor 2)**, que busca reducir drásticamente el tamaño de las bases de datos genómicas sin perder información. Este trabajo forma parte del esfuerzo por hacer que la bioinformática sea más sostenible, accesible y útil para la investigación biomédica.

2. Resumen del artículo

El artículo explica cómo la secuenciación del ADN se ha vuelto mucho más barata y rápida, lo que ha permitido crear grandes proyectos que estudian miles de genomas humanos. Sin embargo, esto genera una enorme cantidad de datos que deben almacenarse y transmitirse.

Para resolver este problema, los autores desarrollaron una herramienta informática llamada **GDC 2 (Genome Differential Compressor 2)**, capaz de **comprimir los genomas humanos miles de veces** sin perder información. Por ejemplo, una colección de 1.092 genomas que normalmente ocuparía 6,7 terabytes puede reducirse a solo 700 megabytes.

El algoritmo no solo es muy eficiente, sino también rápido, procesando los datos a unos 200 MB por segundo. Gracias a este avance, la investigación genética y la medicina personalizada pueden manejar enormes volúmenes de información de forma más económica y accesible.

3. Cuestión 1: Impacto del avance en bioinformática y medicina genómica

En 2015, la genómica vivía una revolución tecnológica. Secuenciar un genoma humano, que a comienzos de los años 2000 costaba millones de dólares, ya podía hacerse por menos

de mil. Esto permitió crear grandes proyectos internacionales con miles de genomas, como el 1000 Genomes Project o el UK10K.

Sin embargo, este avance trajo un nuevo problema: el exceso de datos. Cada genoma ocupa varios gigabytes, y almacenar miles de ellos suponía terabytes de información, lo que hacía muy difícil su manejo, transferencia y análisis.

En este contexto, el trabajo de Deorowicz, Danek y Niemiec (2015) fue pionero. Con GDC 2, los autores presentaron un sistema capaz de comprimir colecciones completas de genomas humanos hasta 9.500 veces su tamaño original, sin perder información. Este logro multiplicó por cuatro la eficiencia de los compresores existentes hasta ese momento y marcó un punto de inflexión en la bioinformática del almacenamiento genómico.

El algoritmo aprovechaba la enorme similitud entre los genomas humanos, reduciendo los datos redundantes mediante un proceso de compresión en dos niveles, y además estaba diseñado para funcionar rápido y en paralelo en equipos modernos.

En su momento, este avance fue crucial porque preparó el terreno para la medicina genómica y la secuenciación masiva. A partir de 2015, herramientas como GDC 2 ayudaron a hacer posible el almacenamiento eficiente de grandes bases de datos de ADN, algo esencial para la investigación en cáncer, enfermedades raras y medicina personalizada.

4. Cuestión 2: Desafíos éticos, técnicos y sociales

En 2015, la secuenciación de ADN se había vuelto tan accesible que comenzaron a acumularse miles de genomas personales en bases de datos públicas y privadas. Esto planteó problemas de privacidad y consentimiento: ¿quién controla esos datos?, ¿cómo evitar que se identifique a una persona por su ADN? También aparecieron dilemas sobre el uso comercial de la información genética, como su venta a farmacéuticas o aseguradoras.

El mayor reto técnico era la gestión y almacenamiento de volúmenes masivos de datos. La infraestructura computacional no estaba preparada para manejar terabytes o petabytes de información genómica. Además, existía la necesidad de estandarizar los formatos y garantizar que la compresión o el procesamiento de datos no alteraran la información biológica original. La velocidad y eficiencia de los algoritmos, como GDC 2, fueron una respuesta directa a este desafío.

A nivel social, la genómica empezó a mostrar una brecha clara: no todos los países ni laboratorios tenían acceso a estas tecnologías. Esto generaba desigualdad científica y limitaba la participación global en proyectos de investigación. También surgió la preocupación por la alfabetización genética, es decir, que la población comprendiera los riesgos y beneficios de compartir su información genética.

5. Referencias

- Deorowicz, S., Danek, A., & Niemiec, M. (2015). *GDC 2: Compression of large collections of genomes*. Scientific Reports, 5:11565. DOI: 10.1038/srep11565. Disponible en: https://www.nature.com/articles/srep11565

- The 1000 Genomes Project Consortium. (2012). *An integrated map of genetic variation from 1092 human genomes*. Nature, 491, 56–65.