

Algoritmos de Brujin

Gisela Belmonte Cruz Diego Marrero Ferrera

Octubre 2025

Resumen

El presente trabajo tuvo como objetivo comprender el proceso de **ensamblaje de secuencias de ADN** a partir de fragmentos cortos mediante los **algoritmos de Brujin**, una herramienta fundamental en bioinformática y en los métodos modernos de secuenciación genómica. A través de tres ejercicios prácticos se abordaron los siguientes aspectos:

- En la **Cuestión 1**, se partió de un conjunto de fragmentos de ADN y se construyó el grafo de Brujin correspondiente a partir de sus prefijos y sufijos de longitud $k - 1$. Con este modelo se observó cómo las lecturas se conectaban mediante aristas que representaban las superposiciones entre fragmentos.
- En la **Cuestión 2**, se repitió el proceso de construcción del grafo con nuevas lecturas para obtener un **camino euleriano**, lo que permitió **reconstruir la secuencia original del ADN**. Se demostró que la estructura del grafo facilitaba el ensamblaje al identificar los solapamientos de manera eficiente.
- En la **Cuestión 3**, se analizó un caso más complejo en el que aparecieron **repeticiones y ambigüedades**, lo que impidió encontrar un camino euleriano único. Se examinaron las causas del problema y se plantearon posibles soluciones, como ajustar la longitud de k o emplear estrategias adicionales para resolver los conflictos de ensamblaje.

En conjunto, los ejercicios permitieron comprender tanto la potencia como las limitaciones de los algoritmos de Brujin en el ensamblaje de genomas, así como su relevancia en los sistemas de secuenciación de próxima generación (NGS).

1. Ejercicios

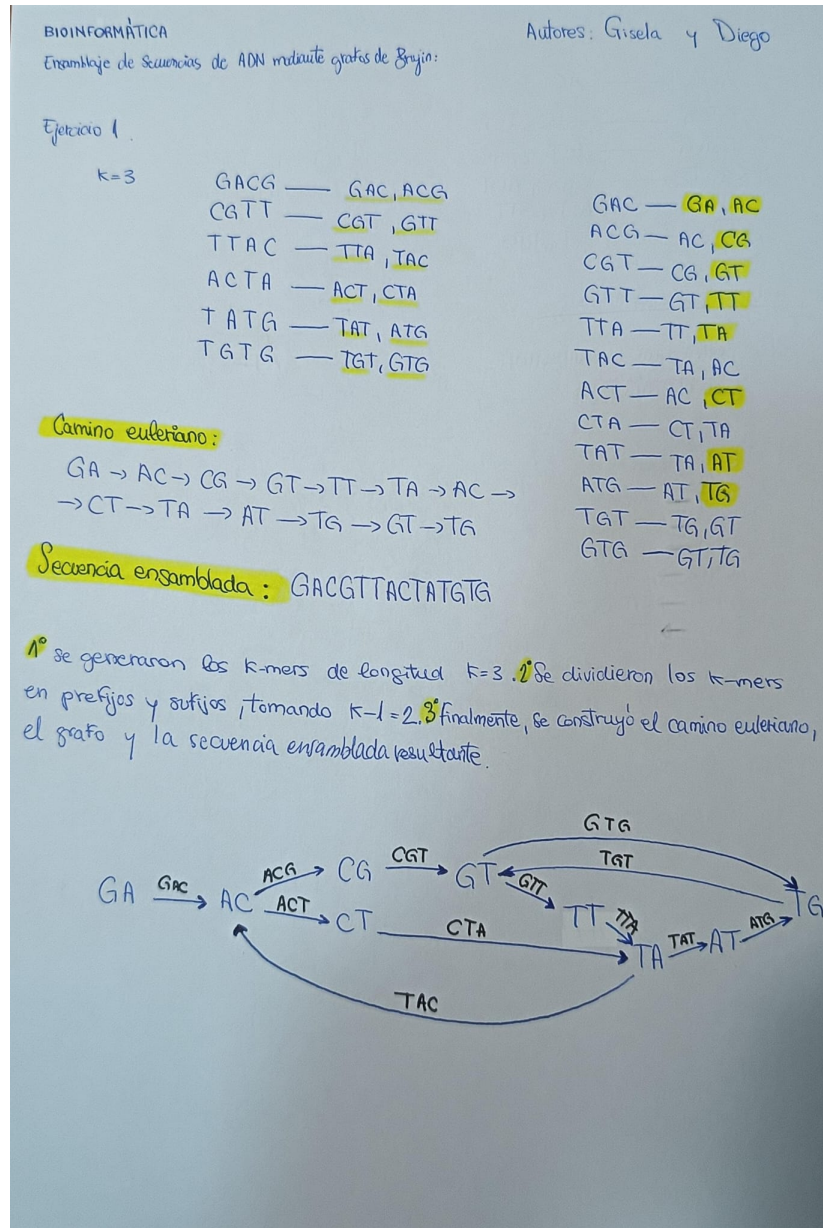


Figura 1: Cuestión 1: Fragmentos iniciales y generación de prefijos y sufijos.

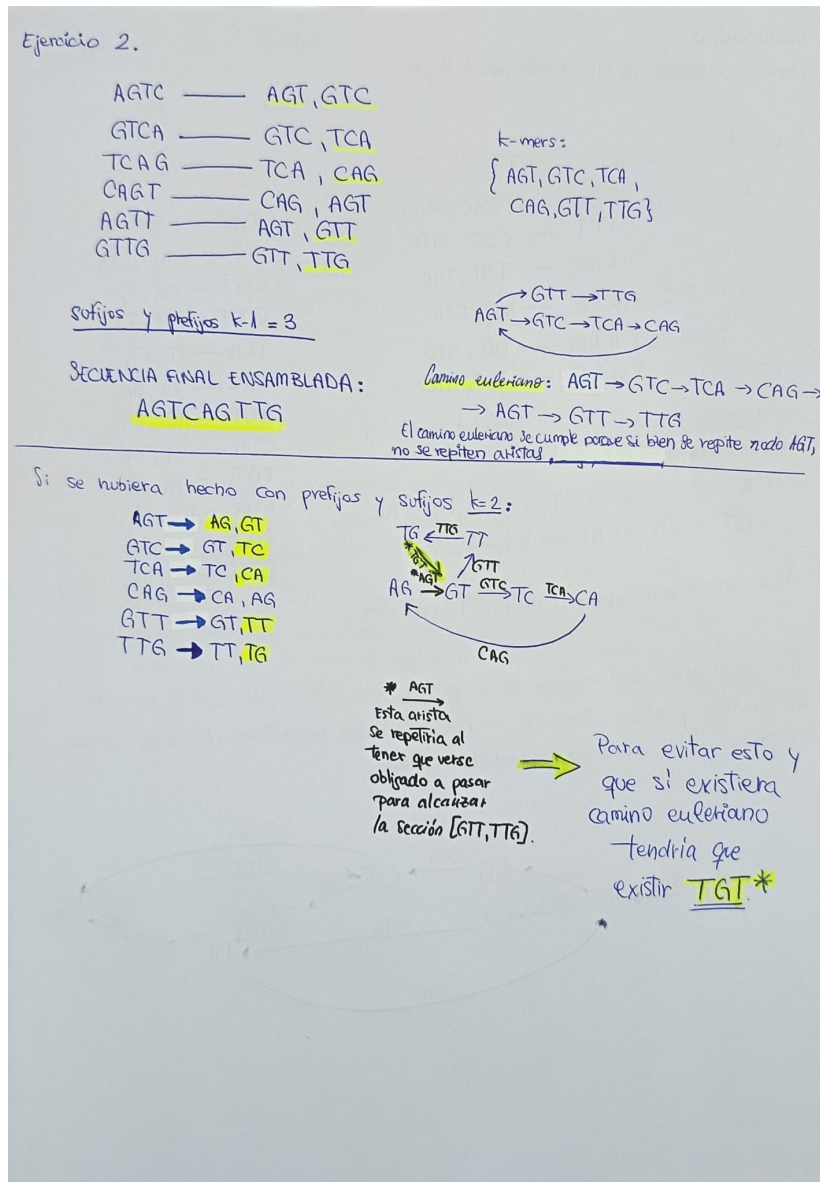


Figura 2: Cuestión 2: Construcción del grafo de Brujin y obtención del camino euleriano.

* Cuestión 3

1. Fragmentos: 2. Aristas del grafo (prefijo \rightarrow sufijo con $k-1=3$)

AGTT \rightarrow AGT \rightarrow GTT
 GTTG \rightarrow GTT \rightarrow TTG
 TTGA \rightarrow TTG \rightarrow TGA
 TGAT \rightarrow TGA \rightarrow GAT
 GACG \rightarrow GAC \rightarrow ACG
 ACGA \rightarrow ACG \rightarrow CGA
 CGAA \rightarrow CGA \rightarrow GAA
 GAAA \rightarrow GAA \rightarrow AAA
 AACG \rightarrow AAC \rightarrow ACG

3. Conjunto de nodos (3-meros) y grados in/out

El conjunto de nodos es
 $\{AGT, GTT, TTG, TGA, GAT, GAC, ACG, CGA, GAA, AAA, AAC\}$

Basándonos en las aristas que tratamos previamente, vemos que hay 3 posibles inicios por desbalances de "out-in", e igual con los finales.

+ ¿Inicios?: AGT: +1; AAC: +1; GAC: +1
 + ¿Finales?: GAT: -1; AAA: -1; ACG: -1

Esto demuestra que NO se puede tratar un camino euleriano único que recorra todas las aristas. Lo máximo que podemos tratar sería:

4. Intentos de ensamblaje (separado).

① AGT \rightarrow GTT
 \downarrow
 TTG
 \downarrow
 GAT \leftarrow TGA

Seq. recons.: AGTTTGAT

② GAC \rightarrow ACG \rightarrow CGA \rightarrow GAA \rightarrow AAA
 AAC \rightarrow ACG

Secuencia 1: GACGAAA

Arista suelta AAC \rightarrow ACG = AACG

5. ¿Por qué no se puede formar en una sola secuencia?

Hay ramificaciones y desbalances entre entradas y salidas (más de un inicio y un final)

Figura 3: Cuestión 3: Detección de ambigüedades y limitaciones del grafo de Brujin.