

Índice

Introducción

Problema

Algoritmos

Función de Fitness

Parámetros

Resultados

Conclusiones

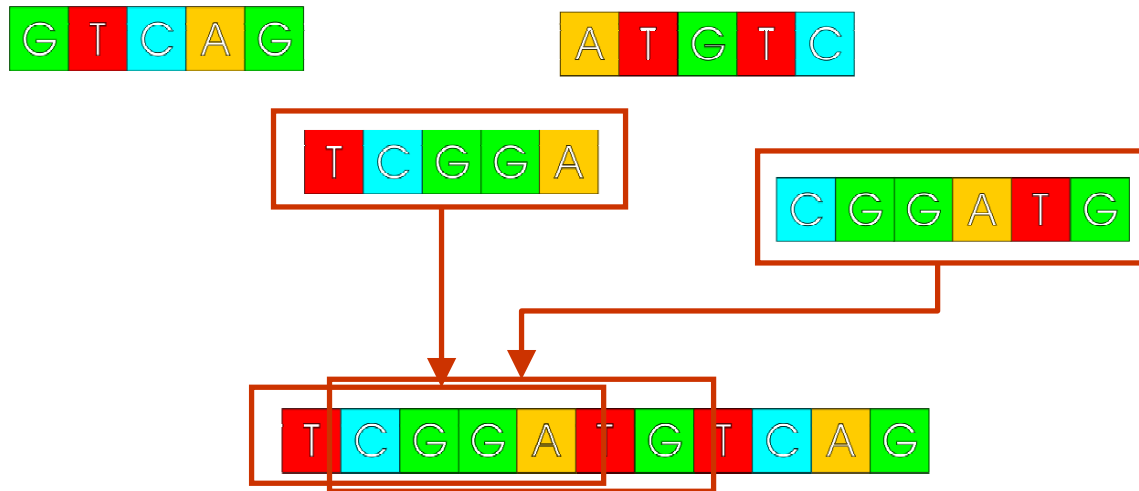
**Ensamblado de Fragmentos de ADN usando
Metaheurísticas Secuenciales y Paralelas**

Universidad de Málaga, ESPAÑA

por **Gabriel Luque**, Enrique Alba, Sami Khuri

Introducción

- El problema consiste en construir una secuencia de ADN a partir de varios fragmentos.



- Métodos de ensamblaje:
 - “A mano”: múltiples errores y pequeñas instancias.
 - Alg. Secuenciales (PHRAP, CAP3, ...): < 34k.
 - Propuesta: algoritmo genético distribuido paralelo.

RADI-AEB'04

Índice

Introducción

→ Problema

Algoritmos

Función de Fitness

Parámetros

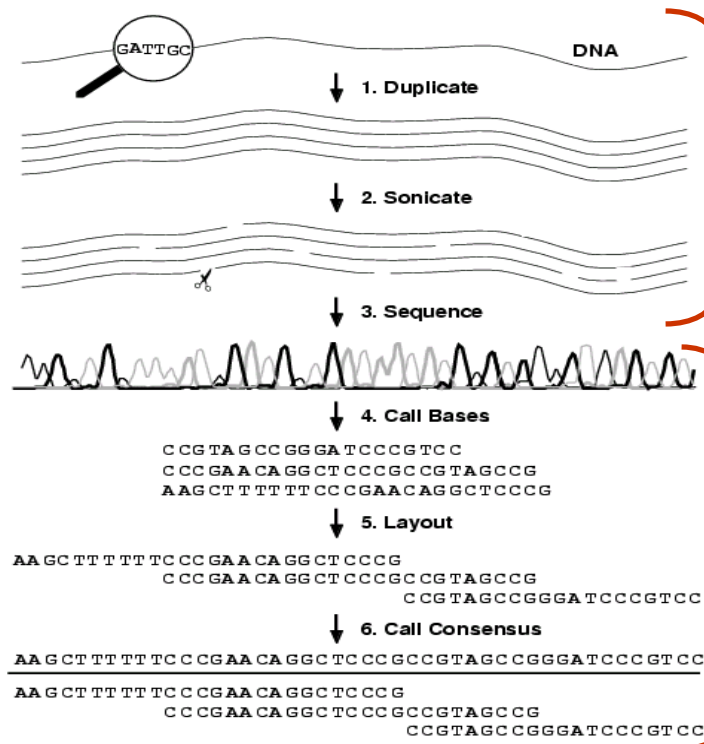
Resultados

Conclusiones

Problema (I)

□ Dado un conjunto de fragmentos de ADN, conseguir un orden que permita reconstruir la cadena original de ADN.

□ Técnica empleada:



Laboratorio

Problema a resolver

RADI-AEB'04

Índice

Introducción

➔ Problema

Algoritmos

Función de Fitness

Parámetros

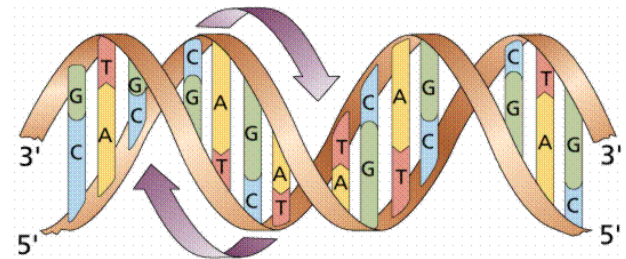
Resultados

Conclusiones

Problema (II)

- ❑ Paso 4: Calcular el solapamiento entre fragmentos:
 - Normalmente se emplea la programación dinámica.

- ❑ Paso 5: Encontrar el orden de los fragmentos:
 - Este es el problema abierto que queremos resolver.
 - Existen varias situaciones que dificultan el orden:
 1. Orientación desconocida.
 2. Error en las bases.
 3. Cobertura incompleta.
 4. Regiones repetidas.



RADI-AEB'04

Índice

Introducción

➔ Problema

Algoritmos

Función de Fitness

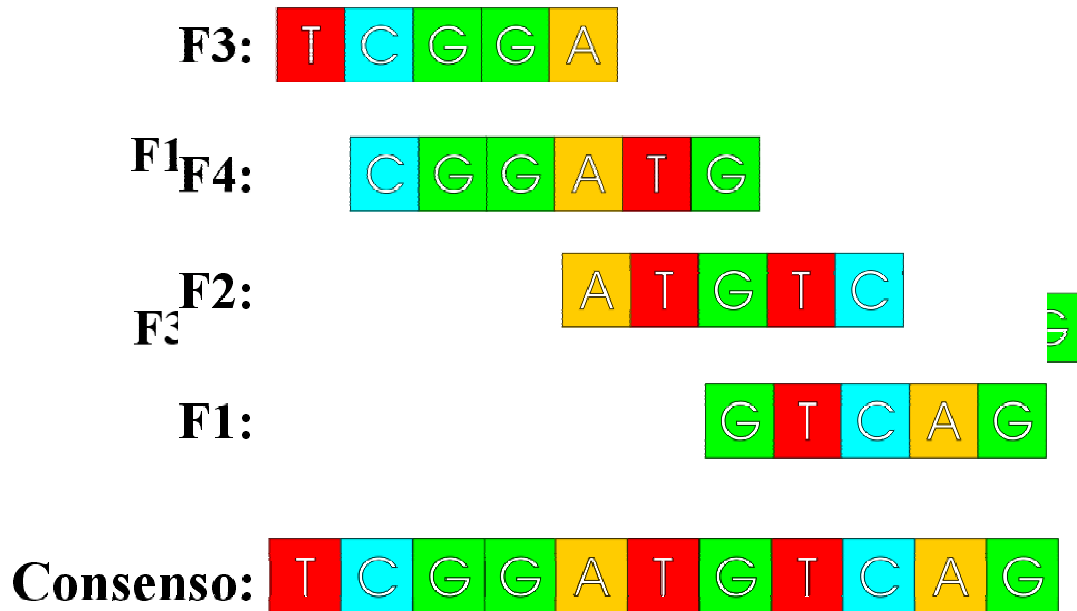
Parámetros

Resultados

Conclusiones

Problema (III)

- ❑ Paso 6: Calcular secuencia final:
 - Regla de consenso para unir fragmentos.
- ❑ Ejemplo:



RADI-AEB'04

Índice

Introducción

Problema

➔ Algoritmos

Función de Fitness

Parámetros

Resultados

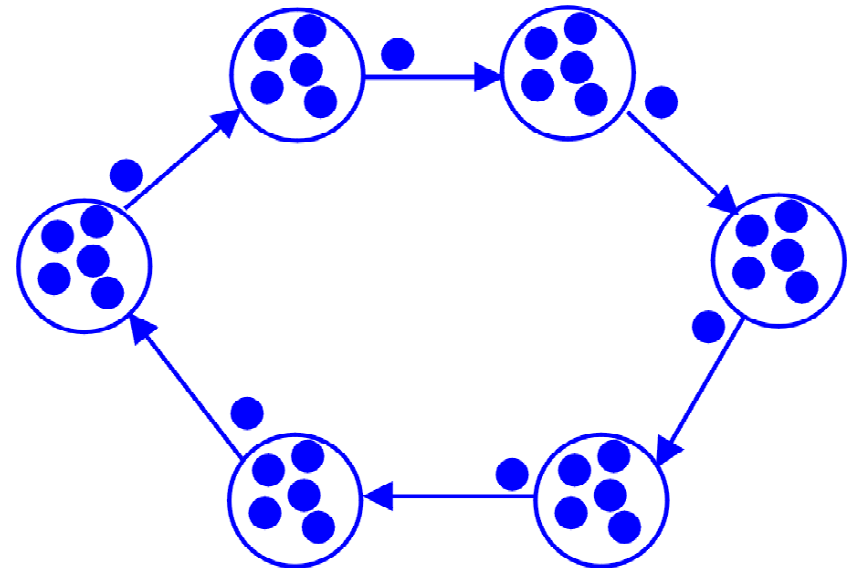
Conclusiones

Algoritmos Utilizados

➤ Algoritmo genético secuencial.

➤ Algoritmo genético distribuido.

- 2, 4, y 8 islas
- Popsizet / n_islas



RADI-AEB'04

Índice

Introducción

Problema

Algoritmos

➔ Función de Fitness

Parámetros

Resultados

Conclusiones

Algoritmo: Función de Fitness

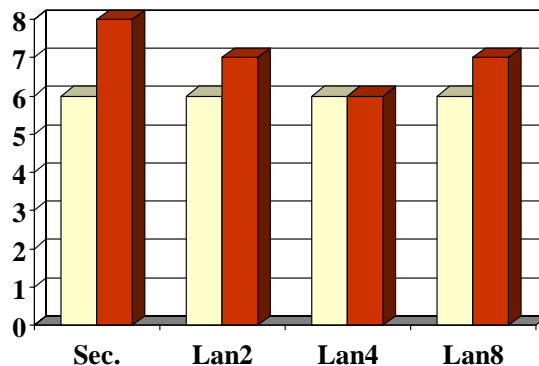
➤ **F1**: Maximiza solapamiento de frags. consecutivos.

$$F1(l) = \sum_{i=0}^{n-2} w(f[i]f[i+1])$$

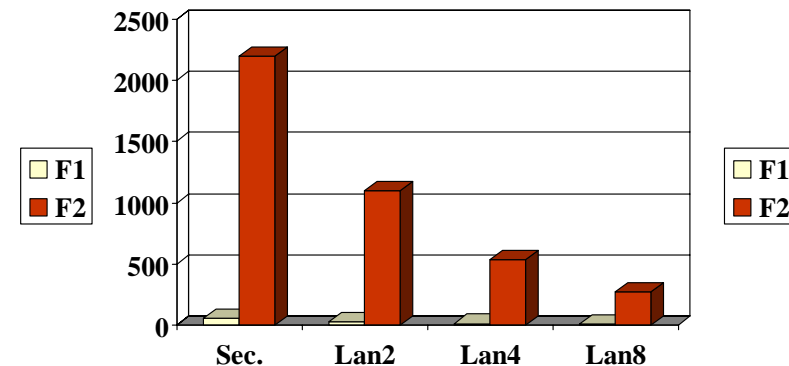
➤ **F2**: Minimiza solapamiento entre frag. lejanos.

$$F2(l) = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} |i-j| \times w(f[i]f[j])$$

Número de Contig



Tiempo (segundos)



RADI-AEB'04

Índice

Introducción

Problema

Algoritmos

Función de Fitness

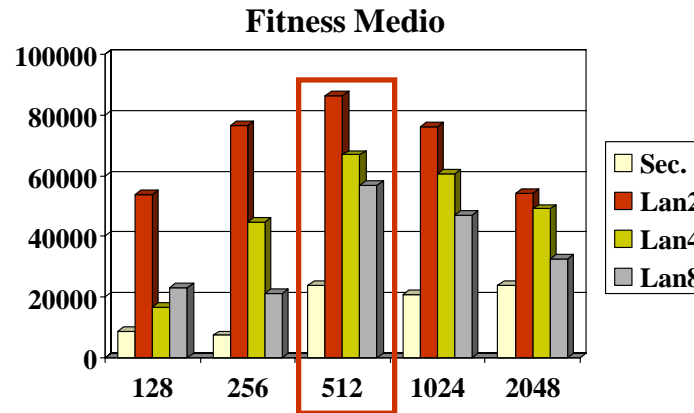
➡ Parámetros

Resultados

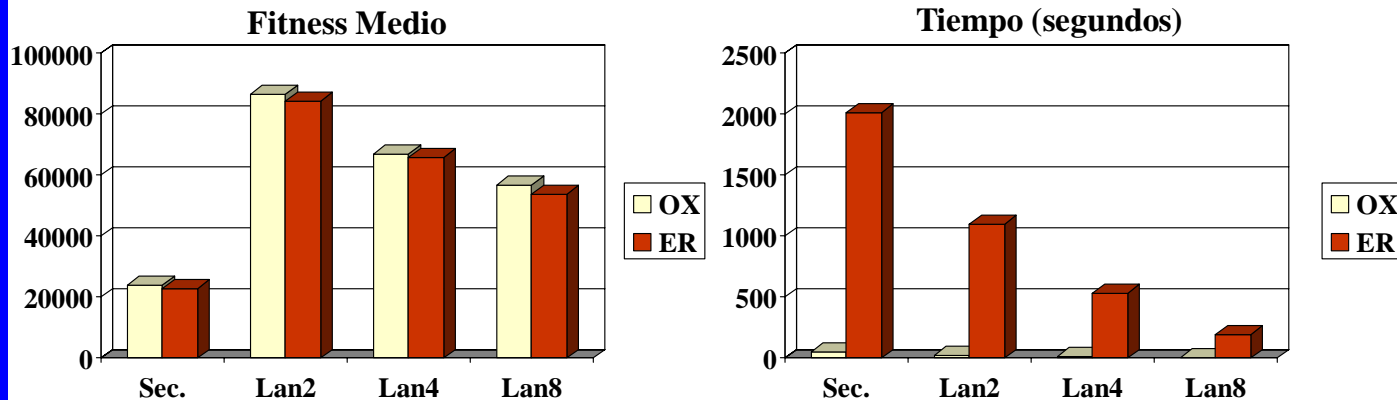
Conclusiones

Algoritmos: Parámetros I

➤ Tamaño de la población: 128, 256, 512, 1024, 2048.



➤ Operador de Cruce: OX y ER.



RADI-AEB'04

Índice

Introducción

Problema

Algoritmos

Función de Fitness

➡ Parámetros

Resultados

Conclusiones

Algoritmos: Parámetros II

➤ Probabilidad de Aplicación de los Operadores:

➤ Cruce: 0.3, 0.5, 0.7, 1.0 ⇒ 1.0

➤ Mutación: 0.0, 0.1, 0.2, 0.3 ⇒ 0.3

➤ Política de Migración:

➤ Frecuencia: 5, 20, 50 ⇒ 20

➤ N° de emigrantes: 1, 10, 20 ⇒ 1

La selección de probabilidades se ha hecho después de Experimentación exhaustiva para todos los valores mostrados

RADI-AEB'04

Índice

Introducción

Problema

Algoritmos

Función de Fitness

Parámetros

→ Resultados

Conclusiones

Resultados: Instancias

- ❑ Secuencia: BX842596 (GI 38524243):
 - Moho del pan.
 - 77.292 bases.

- ❑ Instancia 1: 38524243_4 (Generado con GenFrag):
 - 442 fragmentos.
 - 708 bases de longitud media.
 - Cobertura: 4.

- ❑ Instancia 2: 38524243_7 (Generado con GenFrag):
 - 733 fragmentos.
 - 703 bases de longitud media.
 - Cobertura: 7.

RADI-AEB'04

Índice

Introducción

Problema

Algoritmos

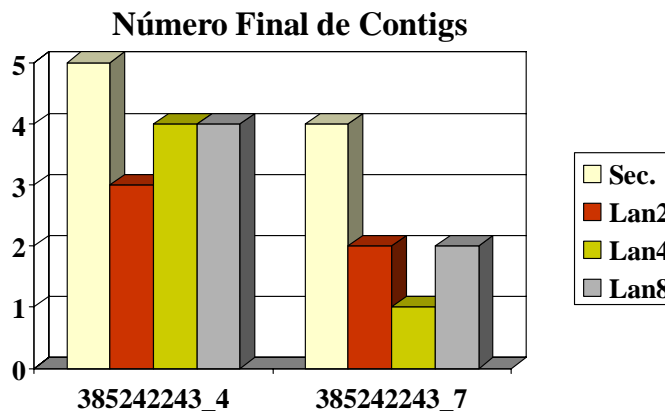
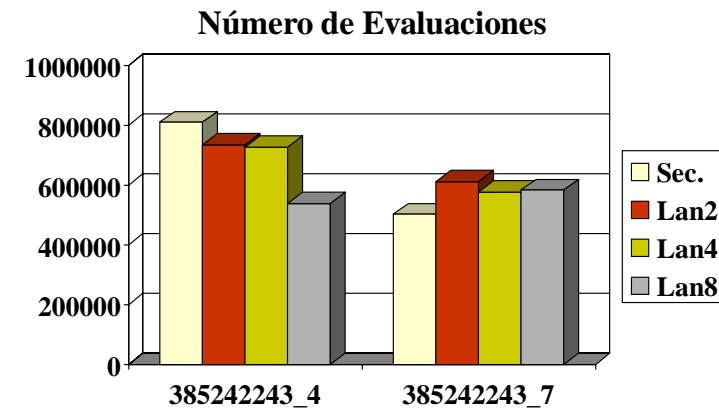
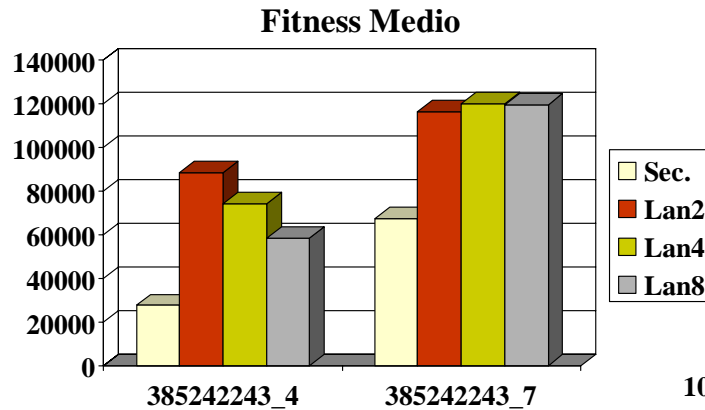
Función de Fitness

Parámetros

→ Resultados

Conclusiones

Resultados (I)



RADI-AEB'04

Índice

Introducción

Problema

Algoritmos

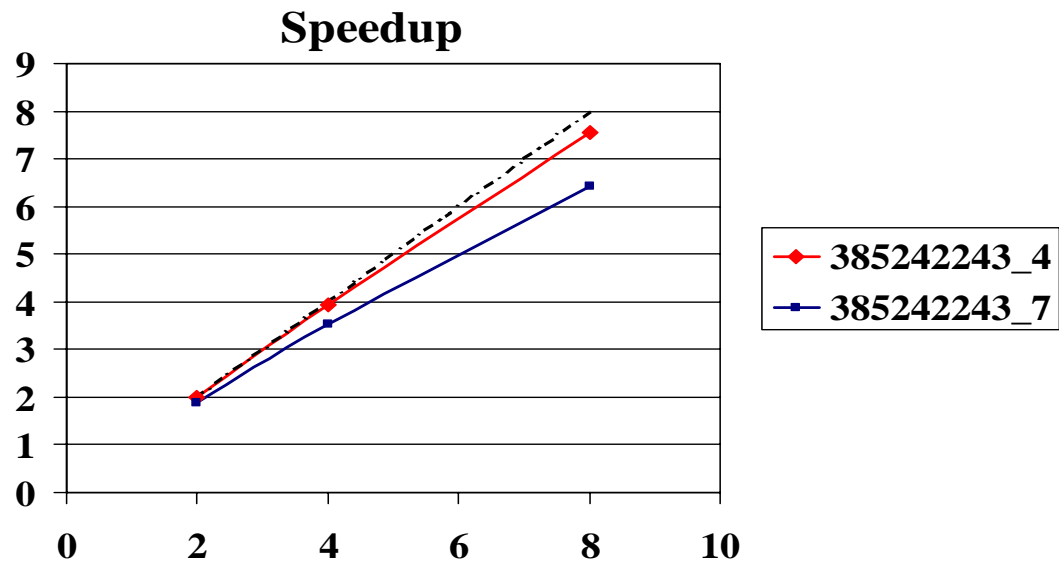
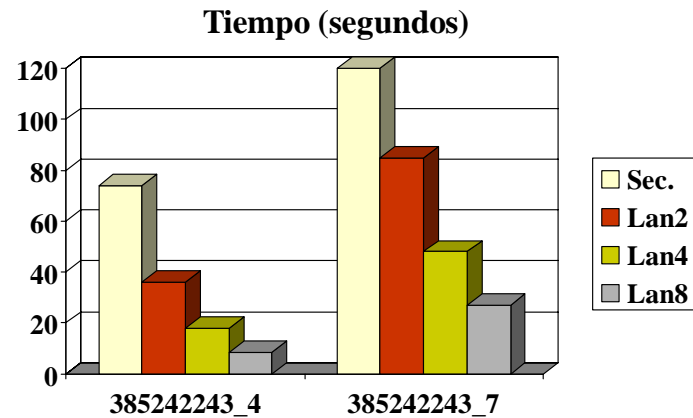
Función de Fitness

Parámetros

 Resultados

Conclusiones

Resultados (II)



RADI-AEB'04

Índice

Introducción

Problema

Algoritmos

Función de Fitness

Parámetros

Resultados

→ Conclusiones

Conclusiones

- ❑ El ensamblado de fragmentos de ADN es un problema **NP** y usualmente es tratado con **heurísticos**.
- ❑ Los **heurísticos secuenciales** tratan **bien** el problema para tamaños **medios** (14-30k) pero para tamaños **grandes** (77k) tienen muchas **dificultades**.
- ❑ El **paralelismo** no solo permite **reducir el tiempo** sino que consigue **resultados más precisos**, incluso llegando a la solución óptima (1 contig).
- ❑ Varias líneas siguen abiertas, como el **desarrollo** de una **función de fitness** más adecuada.

RADI-AEB'04

Índice

Introducción

Problema

Algoritmos

Función de Fitness

Parámetros

Resultados

Conclusiones

¿Preguntas?

Málaga

