**UNIVERSIDAD AUTÓNOMA DE COAHUILA**

**FACULTAD DE SISTEMAS**

**ANÁLISIS COMPARATIVO DE TRES METAHEURÍSTICAS DE OPTIMIZACIÓN PARA EL ALINEAMIENTO DE SECUENCIAS DE ADN DE VIRUS GENÉTICAMENTE SIMILARES**

**INGENIERÍA EN SISTEMAS COMPUTACIONALES**

**FRANCISCO ALAN MENCHACA MERINO**

**PROF. ERNESTO LIÑAN GARCIA**

**SALTILLO, COAH.**  **ABRIL, 2024**

**TABLA DE CONTENIDOS**

[INTRODUCCIÓN 5](#_Toc163646488)

[CAPITULO I. ALINEAMIENTO DE SECUENCIAS ADN 6](#_Toc163646489)

[1. Contextualización del Problema 7](#_Toc163646490)

[2. Fundamentos del Alineamiento de Secuencias 8](#_Toc163646491)

[3. Alineamiento de Secuencias con Algoritmos Genéticos 8](#_Toc163646492)

[3.1 Exploración Eficiente del Espacio de Soluciones 9](#_Toc163646493)

[3.2 Adaptabilidad a la Diversidad Genómica 9](#_Toc163646494)

[3.3 Soluciones Óptimas y Subóptimas 9](#_Toc163646495)

[3.4 Escalabilidad y Paralelismo 9](#_Toc163646496)

[CAPITULO II. METAHEURÍSTICAS 10](#_Toc163646497)

[1. Algoritmo Genético 12](#_Toc163646498)

[Introducción 12](#_Toc163646499)

[Funcionamiento del Algoritmo 12](#_Toc163646500)

[Pseudocódigo del Algoritmo Genético 13](#_Toc163646501)

[2. Algoritmo de Reacciones Químicas 15](#_Toc163646502)

[Introducción 15](#_Toc163646503)

[Estructura Molecular 15](#_Toc163646504)

[Energía Potencial 15](#_Toc163646505)

[Energía Cinética 16](#_Toc163646506)

[Cambio de Estructura Molecular 16](#_Toc163646507)

[Búsqueda de estructura más favorable 16](#_Toc163646508)

[Reacciones elementales en ARC 16](#_Toc163646509)

[Colisiones Ineficaces 17](#_Toc163646510)

[Colisiones Eficaces 17](#_Toc163646511)

[Colisión Ineficaz contra la Pared 17](#_Toc163646512)

[Descomposición 18](#_Toc163646513)

[Colisión Intermolecular Ineficaz 20](#_Toc163646514)

[Síntesis 20](#_Toc163646515)

[3. Algoritmo de Recocido Simulado 21](#_Toc163646516)

[Introducción 21](#_Toc163646517)

[Funcionamiento del Algoritmo 21](#_Toc163646518)

[Pseudocódigo del Recocido Simulado 22](#_Toc163646519)

[CAPITULO III. IMPLEMENTACIÓN DE LOS ALGORITMOS 24](#_Toc163646520)

[Mutador de Secuencias Simple 25](#_Toc163646521)

[Algoritmo Genético 27](#_Toc163646522)

[Modificador Cruce de Secuencias 27](#_Toc163646523)

[Algoritmo de Reacciones Químicas 30](#_Toc163646524)

[Modificador de Reacciones Químicas 30](#_Toc163646525)

[Algoritmo de Recocido Simulado 31](#_Toc163646526)

[Recocido Simulado adaptado al Alineamiento de Secuencias 32](#_Toc163646527)

[CAPITULO IV. EXPERIMENTACIÓN Y RESULTADOS 33](#_Toc163646528)

[Comparativa de Metaheurísticas 34](#_Toc163646529)

[Generación de Secuencias 34](#_Toc163646530)

[Modificadores de Secuencias 34](#_Toc163646531)

[Representación Gráfica y Numérica de Resultados Obtenidos 34](#_Toc163646532)

[Resultados Numéricos 35](#_Toc163646533)

[Media Aritmética y Desviación Típica de Aptitud 35](#_Toc163646534)

[Análisis del Algoritmo de Recocido Simulado 36](#_Toc163646535)

[Parámetros del Algoritmo 36](#_Toc163646536)

[Resultados y Comparativa 36](#_Toc163646537)

[REFERENCIAS BIBLIOGRAFICAS 38](#_Toc163646538)

# INTRODUCCIÓN

La comprensión del funcionamiento y la estructura del ADN es un campo fundamental en la biología y la genética [1]. En el análisis genético, el estudio del ADN se ha vuelto indispensable para revelar la evolución, comportamiento y cambios (mutaciones) en los genes que pueden ocasionar posibles enfermedades o afecciones.

La Alineación de Múltiples Secuencias (AMS) de ADN desarrolla un función esencial en la comparación de secuencias genéticas para identificar mutaciones, regiones conservadas y otros patrones o tendencias importantes [2], facilitando varios análisis biológicos importantes, incluida la predicción de estructuras proteicas desconocidas, esta tarea es desafiante a la complejidad de las secuencias y la diversidad de procesos biológicos que pueden afectar en su evolución, en este contexto, la identificación de subsecuencias máximamente homólogas entre conjuntos de largas secuencias son un problema importante en el análisis de secuencias genéticas.

Para encontrar soluciones aproximadas en problemas como el ASM [3] los algoritmos bio-inspirados han emergido como valiosas herramientas computacionales. Estas técnicas se inspiran en heurísticas inteligentes o procesos naturales para explorar el espacio de soluciones y encontrar configuraciones que satisfagan ciertas métricas objetivo.

Si bien existen numerosos métodos disponibles para AMS, los algoritmos bio-inspirados destacan por su eficiencia. A pesar del creciente interés de la investigación por abordar el desafío del AMS, sólo un puñado ha llevado revisiones exhaustivas en este ámbito. Para cerrar esta brecha, este estudio lleva un análisis de tres metaheurísticas (algoritmos) adaptas para la solución de este tipo de problemas.

Las metaheurísticas a analizar son el Algoritmo Genético (AG), el Algoritmo de Recocido Simulado (ARS) y el Algoritmo de Reacciones Químicas (ARQ), todas ellas adaptadas para resolver el problema del AMS.

# CAPITULO I. ALINEAMIENTO DE SECUENCIAS ADN

## Contextualización del Problema

A medida que avanzamos en el siglo XXI, el acceso a datos genómicos se ha vuelto más accesible y abundante que nunca. La era de la secuenciación masiva del ADN ha generado mucha información genómica proveniente de diversas especies y fuentes.

A pesar de ser un logro destacado, esta abundancia de datos plantea desafíos significativos que requieren abordajes innovadores. Uno de los problemas más destacados en este contexto es el alineamiento de secuencias de ADN. La comparación de secuencias de ADN es esencial para identificar similitudes y diferencias, y para comprender la evolución y la función de las secuencias genéticas.

Con el crecimiento del volumen de datos genómico, se presentan desafíos considerables. En primer lugar, las secuencias de ADN son inherentemente largas y complejas, y la variabilidad en las secuencias puede dificultar la identificación de regiones conservadas. Además, las secuencias pueden contener inserciones, deleciones, y mutaciones que complican aún más la tarea de alinearlas de manera precisa.

La identificación de similitudes puede requerir el procesamiento de miles o incluso millones de secuencias, lo que plantea un desafío de escalabilidad en términos de tiempo y recursos computacionales.

El alineamiento de secuencias se utiliza en una amplia gama de aplicaciones, desde la identificación de genes homólogos en diferentes especies hasta el estudio de mutaciones genéticas relacionadas con enfermedades.

En la investigación biomédica, el alineamiento de secuencias de ADN es esencial para comprender la genética de poblaciones, la evolución molecular y la genómica comparativa. Por lo tanto, es crucial desarrollar enfoques que sean capaces de abordar de manera eficiente y precisa el alineamiento de secuencias de ADN en este contexto de crecimiento exponencial de datos genómicos.

Los enfoques tradicionales, como los algoritmos de programación dinámica, pueden volverse prohibitivamente costosos computacionalmente en este entorno. Esto motiva la búsqueda de soluciones innovadoras, como la aplicación de Algoritmos Genéticos, que tienen el potencial de manejar datos masivos y resolver los desafíos planteados por el alineamiento de secuencias de ADN en la era de la genómica de próxima generación. En los capítulos subsiguientes, exploraremos en detalle cómo los Algoritmos Genéticos pueden contribuir a la solución de estos problemas y proporcionar una nueva perspectiva en el alineamiento de secuencias de ADN.

## Fundamentos del Alineamiento de Secuencias

Los avances tecnológicos en secuenciación de ADN han acelerado la generación de datos genómicos a una escala sin precedentes, se ha vuelto evidente que el alineamiento de secuencias desempeña un papel central en la extracción de información biológica y la comprensión de la genética.

El ADN, como la molécula de la vida, alberga la información genética que rige la formación y funcionamiento de todos los seres vivos. La diversidad en las secuencias de ADN, que puede manifestarse a través de variaciones genéticas, mutaciones y la presencia de elementos genéticos móviles, ha llevado a la aparición de nuevos desafíos. El alineamiento de secuencias de ADN se convierte, en consecuencia, en un pilar esencial para abordar cuestiones clave en la genómica y la biología molecular.

En este contexto, el alineamiento de secuencias desempeña un papel fundamental en la investigación genómica, la identificación de genes, la comprensión de la evolución y la medicina personalizada. La explosión de datos genómicos, impulsada por la secuenciación de nueva generación, ha amplificado la importancia de desarrollar enfoques innovadores para abordar los desafíos que plantea la gestión y el análisis de datos masivos.

## Alineamiento de Secuencias con Algoritmos Genéticos

Los AG (Algoritmos Genéticos) representan una prometedora revolución en la forma en que abordamos el desafiante campo del alineamiento de secuencias. Al igual que la evolución biológica impulsa la diversidad y la adaptación en la naturaleza, los AG han demostrado su capacidad para explorar soluciones innovadoras y eficientes en el alineamiento de secuencias de ADN, ARN y proteínas.

La promesa de los Algoritmos Genéticos radica en su inspiración en la evolución, lo que les permite abordar cuestiones complejas con un enfoque basado en poblaciones, selección natural y reproducción. Este enfoque único trae consigo una serie de ventajas clave:

### Exploración Eficiente del Espacio de Soluciones

Los AG pueden explorar un amplio espacio de soluciones de alineamiento en paralelo, lo que los convierte en herramientas ideales para abordar la diversidad genómica. Múltiples secuencias y posibles soluciones se someten a procesos de cruces genéticos y mutaciones, lo que permite la generación de una variedad de soluciones en un solo proceso.

### Adaptabilidad a la Diversidad Genómica

La adaptabilidad es un elemento fundamental en la evolución, y los AG incorporan esta característica al alineamiento de secuencias. Pueden lidiar con la variabilidad en la longitud de las secuencias, la presencia de inserciones y deleciones, así como mutaciones, sin requerir una alineación predefinida.

### Soluciones Óptimas y Subóptimas

Los AG no solo se centran en encontrar soluciones óptimas, sino que también exploran soluciones subóptimas. Esto es valioso cuando se trabaja con secuencias con múltiples soluciones válidas. Los AG permiten evaluar y seleccionar soluciones en función de su aptitud, lo que proporciona flexibilidad en la elección de resultados.

### Escalabilidad y Paralelismo

En un contexto de crecimiento exponencial de datos genómicos, la escalabilidad y el paralelismo son fundamentales. Los AG pueden implementarse en sistemas de alto rendimiento para manejar grandes volúmenes de secuencias, acelerando significativamente el alineamiento. La aplicación de Algoritmos Genéticos al alineamiento de secuencias representa un emocionante avance que promete abordar los desafíos y las oportunidades que se presentan en la genómica contemporánea.

# CAPITULO II. METAHEURÍSTICAS

El AMS suele ser un verdadero desafío, ya que comúnmente se suele trabajar con “” número de secuencias, lo que hace que encontrar los mejores alineamientos en múltiples secuencias sea un tarea complicada, además, frecuentemente se generan poblaciones de “” número de secuencias mutadas por cada secuencia, es decir, número de secuencias totales, en consecuencia, resolver el AMS es un problema que también involucra tener suficiente recursos computacionales (esto según el número de secuencias con las que se trabaje).

Teniendo en cuenta lo previamente dicho, para reducir la complejidad y reducir costos computacionales al tratar de abordar el problema del AMS, al mismo tiempo que se realiza la comparativa de los tres algoritmos bio-inspirados, la cantidad de secuencias con las que se trabaja a lo largo de este proyecto se reducen a solo 2 secuencias, el genoma del Virus de la Inmunodeficiencia Humana (VIH) de un homo sapiens (un ser humano moderno) y un genoma del mismo virus pero de un simio, limitar el estudio a estas dos secuencias, nos ayuda a mantener un enfoque claro y evitar la dispersión en comparaciones más amplias o menos relevantes y a enfocarnos a la evaluación de la eficacia de los algoritmos bio-inspirados.

La secuencia del VIH del simio es la es la secuencia seleccionada para generar poblaciones de “” número de secuencias con gaps (mutaciones) al implementar el Algoritmo Genético y el Algoritmo de Reacciones Químicas, y es el genoma que sufre mutaciones al aplicar el Algoritmo de Recocido Simulado y el Algoritmo de Reacciones Químicas.

Los algoritmos y las nuevas secuencias mutadas generadas a partir de implementar cada uno de los algoritmos son comparados por medio de su “grado de similitud” o simplemente su “similaridad”, la cual es una variable cuantitativa que, por cada secuencia con mutaciones, contabiliza el número de alineamientos (gen por gen) con respecto al virus de VIH del homo sapiens, “similaridad” es como se le llama a la métrica para evaluar el rendimiento de las tres metaheurísticas.

Las secuencias con mayor “similaridad”, son consideradas las “mejores secuencias”, por su alto grado de alineamiento con respecto a la secuencia de referencia (virus de VIH del homo sapiens).

## Algoritmo Genético

### Introducción

El alineamiento de secuencias, una disciplina esencial en la bioinformática y la genómica computacional, es fundamental para comprender las relaciones entre las biomoléculas y las similitudes entre ellas. En lugar de depender de enfoques convencionales, este trabajo propone la adaptación de Algoritmos Genéticos (AG) para abordar el alineamiento de secuencias.

El enfoque del Algoritmo Genético se base en utilizar una secuencia de ADN como referencia para generar poblaciones de secuencias con mutaciones, para después, nuevamente, generar poblaciones de secuencias con mutaciones, mediante cruces genéticos, en línea con el algoritmo clásico del “Kapsack Problem”.

### Funcionamiento del Algoritmo

#### Inicialización de la Población:

El proceso comienza con la secuencia de ADN del virus del VIH del simio como punto de partida. A partir de esta secuencia de genes, se generan múltiples secuencias (“” número de secuencias) con mutaciones aleatorias en sus genes.

La población inicial se compone de estas secuencias mutadas, formando una población diversa de secuencias genéticamente distintas.

#### Agrupación en Pares para Cruce de Genes:

Las secuencias se organizan en pares, como parte de la preparación para la operación de cruce, de modo que “” (la cantidad de secuencias generadas previamente) tiene que ser un número par, si este no es el caso, existirá una secuencia la cual no tendrá una secuencia par para realizar el cruce de genes, para resolver este problema, se toma la secuencia sin par y se selecciona otra secuencia previamente escogida para realizar el último cruce de secuencias.

#### Generación de Punto de Cruce:

En cada par de secuencias, se elige aleatoriamente un punto de cruce. Este punto de cruce determina la posición donde se producirá el intercambio de fragmentos de genes entre las dos secuencias, en un proceso análogo al algoritmo del "Knapsack Problem".

#### Cruce de Genes:

Los genes de las secuencias parentales se cruzan mediante el intercambio de los segmentos de genes a ambos lados del punto de cruce. Esto da lugar a nuevas secuencias que heredan características de ambas secuencias parentales y promueven la diversidad genética en la población.

#### Evaluación del Algoritmo:

Cada secuencia de la población generada a partir del cruce de genes se somete a una comparativa con respecto a la secuencia del virus del VIH del homo sapiens, las secuencias con mayor grado de similitud (con más alineamientos) se consideran las más prometedoras.

#### Condiciones de Parada:

* Un número máximo de generaciones/poblaciones se alcanza.
* Se converge hacia una solución con un grado de similitud aceptable.
* Se alcanza un límite de tiempo.

El algoritmo se repite durante múltiples generaciones o hasta que se cumplan condiciones de parada predefinidas, como un número máximo de generaciones, la convergencia hacia una solución aceptable o un límite de tiempo.

#### Selección del Mejor Genoma de ADN:

Al concluir el proceso, se selecciona la secuencia con el mayor grado de similitud como la solución óptima de alineamiento.

### Pseudocódigo del Algoritmo Genético

1. Inicializar una secuencia de referencia como punto de partida.
2. Generar una población inicial de secuencias mutadas a partir de la secuencia de referencia.
3. Mientras no se cumplan las condiciones de parada:
   1. Organizar las secuencias en pares para el cruce de genes.
   2. Para cada par de secuencias:
      1. Elegir aleatoriamente un punto de cruce.
      2. Realizar el cruce de genes intercambiando fragmentos de genes en ambos lados del punto de cruce.
      3. Agregar las nuevas secuencias generadas a la nueva población.
   3. Retornar la nueva población.
   4. Incrementar el contador de generaciones.
4. Seleccionar la secuencia con el mayor grado de similitud como la solución óptima de alineamiento.

## Algoritmo de Reacciones Químicas

### Introducción

El alineamiento de genes en secuencias es una tarea esencial para descifrar la evolución y las funciones de los genes, y la demanda de algoritmos precisos y eficientes se ha intensificado con la expansión constante de los datos genómicos.

En este contexto, el Algoritmo de Reacciones Químicas presenta una perspectiva innovadora. Este algoritmo se inspira en los principios de la química para abordar el desafío del alineamiento de genes. Este enfoque representa las soluciones como moléculas, donde la "energía" de estas moléculas refleja la calidad de los alineamientos (su similaridad). La "energía cinética" permite a estas moléculas buscar soluciones de mayor calidad.

El algoritmo se basa en reacciones químicas, como colisiones ineficientes, descomposiciones y síntesis, para mejorar los alineamientos de genes. Por ejemplo, las colisiones ineficientes representan soluciones de calidad similar, mientras que las descomposiciones conducen a soluciones radicalmente diferentes.

### Estructura Molecular

Resume las características y sus correspondencias a una solución en el dominio de las matemáticas. Una molécula posee dos tipos de energías:

* Energía Potencial
* Energía Cinética

### Energía Potencial

Considerando que:

* Es una estructura molecular
* Es una función objetivo

La función objetivo para calcular la energía potencial de la estructura molecular se define como:

La función objetivo se encarga de calcular el alineamiento de genes de dos secuencias seleccionadas, el alineamiento obtenido es equivalente a la energía potencial de ambas secuencias.

### Energía Cinética

Representa el cambio en la estructura molecular (de a ). Un cambio es posible si la energía potencial de la nueva estructura es menor o igual que la energía potencial de la estructura original . Si no es así, se permite el cambio solo si .

### Cambio de Estructura Molecular

Una estructura molecular tiene la intención de cambiar de a . El cambio siempre es posible si de lo contrario, permitimos el cambio solo cuando . Cuanto mayor sea la de la molécula, mayor será la posibilidad de que la molécula pueda poseer una nueva estructura molecular con mayor .

### Búsqueda de estructura más favorable

Una estructura más favorable es aquella que tiene el estado potencial más bajo posible.

Las moléculas involucradas en una reacción intentan alcanzar el estado potencial más bajo posible, pero la búsqueda ciega de estructuras más favorables dará como resultado estados metaestables (estructuras menos favorables) es decir, atascarse en mínimos locales.

La (Energía Cinética) permite que las moléculas se muevan a un estado de mayor potencial, y, por lo tanto, la posibilidad de tener una estructura más favorable en un cambio futuro. La Energía Cinética de una molécula simboliza su capacidad de escapar de un mínimo local.

### Reacciones elementales en ARC

* Colisión ineficaz contra la pared
* Descomposición
* Colisión intermolecular ineficaz
* Síntesis

La colisión ineficaz contra la pared y la descomposición son reacciones unimoleculares que se desencadenan cuando la molécula golpea una pared del recipiente. La colisión intermolecular ineficaz y la síntesis involucran a más de una molécula. Ocurren cuando las moléculas chocan entre sí.

### Colisiones Ineficaces

La colisión ineficaz contra la pared y la colisión intermolecular ineficaz reaccionan mucho menos vigorosamente que la descomposición y la síntesis. Las colisiones ineficaces corresponden a aquellos casos en los que las moléculas obtienen nuevas estructuras moleculares en sus propios vecindarios en . Por lo tanto, la de las moléculas resultantes tiende a acercarse a las de las previas.

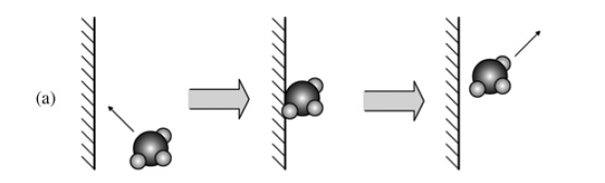
### Colisiones Eficaces

La descomposición y la síntesis tienden a obtener nuevas estructuras moleculares que pueden estar muy lejos de sus vecindarios inmediatos en .

En comparación con las colisiones ineficaces, las colisiones eficaces tienden a que las moléculas resultantes tengan un mayor cambio en que las previas.

### Colisión Ineficaz contra la Pared

Una colisión ineficaz contra la pared se produce cuando una molécula impacta contra la superficie de la pared y luego rebota. Durante esta colisión, algunos atributos moleculares pueden experimentar cambios menores, lo que resulta en una variación limitada en la estructura molecular.



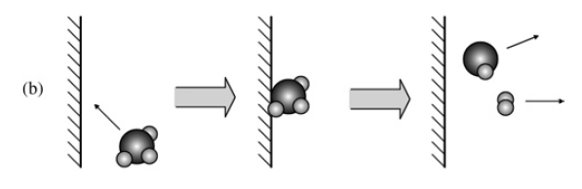
Dado que la colisión carece de una fuerza significativa, la estructura molecular resultante no debería diferir sustancialmente de la estructura original.

Suponiendo que la actual estructura molecular sea , la molécula pretende obtener una nueva estructura en el vecindario de en la (Energía Potencial Superficial) durante la colisión.

La mutación de la molécula es permitida si: Donde:

### Descomposición

La descomposición se produce cuando una molécula choca intensamente contra la pared, lo que resulta en su ruptura en dos o más fragmentos. Esta colisión es vigorosa y provoca una marcada transformación en las estructuras moleculares resultantes, las cuales tienden a ser considerablemente distintas de la estructura original.



Suponiendo que la estructura molecular de la molécula original es y de las moléculas resultantes son y , el cambio es permitido, si la molécula original tiene suficiente energía () para dotar la de las moléculas resultantes, es decir:

Supongamos que:

Donde es un número aleatorio generado uniformemente a partir del intervalo .

Sin embargo, es bastante inusual que la siguiente ecuación se cumpla:

En casos normales, , y tiene valores similares (pero mucho más grandes que los del mismo vecindario), la ecuación solo se cumple cuando es lo suficientemente grande.

Sin embargo, la energía cinética () de las moléculas tiende a disminuir en una secuencia de colisiones ineficaces contra la pared a medida que evoluciona el proceso químico (en consecuencia, la energía cinética no es suficiente).

Por lo tanto, no es probable que la siguiente ecuación se cumpla en casos normales:

Para fomentar la descomposición, utilizamos la energía almacenada en el búfer central para mantener y .

Si esta ecuación se cumple, se permite el cambio y calculamos:

Donde , , y son número aleatorios generados de manera independiente y uniforme a partir del intervalo , la multiplicación por dos números aleatorios asegura que los valores asignados a y no sean demasiado grandes, ya que el búfer suele ser grande.

El búfer se actualiza:

Si las siguientes dos ecuaciones no se cumplen:

Entonces, la descomposición falla y la molécula mantiene su , y originales

### Colisión Intermolecular Ineficaz

Una colisión intermolecular ineficaz describe la situación en la que dos moléculas chocan entre sí y luego rebota el efecto del cambio de energía de las moléculas es similar al de una colisión ineficaz contra la pared, pero esta reacción elemental involucra más de una molécula y no se obtiene Energía Cinética del buffer central.

Suponiendo que las estructuras moleculares originales son y , respectivamente, se acepta los cambios en las moléculas solo si:

### Síntesis

Una síntesis representa más de una molécula (dos moléculas) que chocan y se combinan. Suponiendo que las estructuras moleculares de las dos moléculas original son y , intentamos generar una nueva molécula con estructura molecular a partir de las dos y existentes, dado que la síntesis es vigorosa, debería ser bastante diferente de y , como en la descomposición, se puede utilizar cualquier mecanismo que combine y para formar .

Se acepta solo si:

Retenemos , , , , y , en lugar de , y , si la ecuación no se cumple.

Curiosamente, es grande en comparación con o , porque se espera que tenga un valor similar a o en general.

## Algoritmo de Recocido Simulado

### Introducción

El ARS (Algoritmo de Recocido Simulado) es una técnica de optimización que se inspira en el proceso físico del recocido de metales. Este algoritmo se ha aplicado en una variedad de problemas complejos, desde la optimización de sistemas logísticos hasta el diseño de circuitos electrónicos. Su capacidad para encontrar soluciones de alta calidad en espacios de búsqueda difíciles lo convierte en una herramienta valiosa en la optimización combinatoria.

El ARS es especialmente eficaz en situaciones donde es difícil garantizar la convergencia a una solución óptima. A menudo se utiliza cuando se busca un equilibrio entre exploración y explotación en el espacio de búsqueda, lo que lo hace adecuado para problemas donde existen múltiples soluciones aceptables.

Este algoritmo se basa en la idea de permitir movimientos entre soluciones, incluso si empeoran temporalmente la calidad de la solución, lo que lo hace más flexible que algunos métodos de optimización más rígidos. El ARS puede escapar de mínimos locales y explorar regiones lejanas del espacio de búsqueda.

### Funcionamiento del Algoritmo

A continuación, se presenta un resumen del funcionamiento del Recocido Simulado:

1. Inicialización:El algoritmo comienza utilizando la secuencia de ADN del virus del VIH del simio como solución inicial, esta es la secuencia que sufre mutaciones (generando nuevas secuencias) durante toda la ejecución del algoritmo. Posteriormente se toma como secuencia de referencia el genoma del virus del VIH del homo sapiens, para comparar su grado de similitud con respecto a la secuencia del simio del mismo virus (y para comparar su grado de similitud con las nuevas secuencias generadas del simio).

2. Evaluación: Se calcula el valor de una función objetivo para la solución inicial. Esta función objetivo representa la calidad de la solución, en términos del actual problema a abordar, representa el alineamiento de dos secuencias (su similaridad).

3. Iteración: El Recocido Simulado opera en un bucle iterativo. En cada iteración, se realiza un proceso de perturbación en la solución actual (se añaden mutaciones a la secuencia del vius del VIH del simio). Esta perturbación puede implicar cambios pequeños o grandes en la solución.

4. Evaluación de la solución perturbada: Se calcula el valor de la función objetivo para la solución perturbada. Si la solución perturbada es mejor que la actual, es decir, si tiene un mayor grado de similitud, se acepta automáticamente como la nueva solución actual.

5. Aceptación de soluciones no óptimas: El Recocido Simulado tiene la característica única de que también acepta soluciones que empeoran la calidad de la solución actual. La probabilidad de aceptar una solución no óptima se determina mediante una función que considera la diferencia en la función objetivo y un parámetro denominado "temperatura".

6. Control de temperatura: La temperatura es un parámetro crítico en el Recocido Simulado. Inicialmente, la temperatura es alta, lo que permite la aceptación de soluciones subóptimas con mayor probabilidad. A medida que avanza el proceso, la temperatura se reduce gradualmente, lo que disminuye la probabilidad de aceptar soluciones no óptimas. Este proceso imita el enfriamiento de un metal.

7. Criterio de parada: El algoritmo continúa iterando hasta que se cumple algún criterio de parada. Los criterios comunes incluyen un número máximo de iteraciones, un valor mínimo de temperatura o la convergencia a una solución aceptable.

8. Salida de la mejor secuencia:Una vez que se alcanza el criterio de parada, el Recocido Simulado produce la secuencia con el mayor grado de similitud encontrado durante todo el proceso.

El Recocido Simulado es eficaz para buscar soluciones de alta calidad en espacios de búsqueda complejos y no convexas. Su capacidad para aceptar soluciones no óptimas en las etapas iniciales y su enfoque gradual de disminución de la temperatura lo convierten en una herramienta versátil de optimización.

### Pseudocódigo del Recocido Simulado

#### Variables y Parámetros:

* : Es la probabilidad de aceptar una solución durante el proceso de Recocido Simulado.
* : Un parámetro que controla la probabilidad de aceptar soluciones peores. Generalmente se le conoce como "temperatura".
* : Un contador de iteraciones.
* : Representa la solución actual en la iteración .
* : Es la solución inicial.
* : Parámetro que controla la cantidad de iteraciones en cada etapa del algoritmo.
* : Conjunto de soluciones vecinas a la solución actual .
* : La función objetivo que se está tratando de optimizar, evaluada en la solución .

A white background with black text

Description automatically generated

#### Algoritmo:

1. Inicializa a y a .
2. Repite el proceso hasta que cumpla un criterio de terminación.
3. En cada iteración , realiza lo siguiente:

* Inicializa un bucle para desde hasta .
* Genera una solución vecina seleccionada de .
* Si , entonces acepta la solución como la nueva solución actual (actualiza ).
* Si , calcula la probabilidad de aceptar la solución según la fórmula:

A mathematical equation with black text

Description automatically generated

* Compara , con un valor aleatorio en el rango . Si el valor aleatorio es menor que , acepta la solución como la nueva solución actual (actualiza ).

1. Incrementa en una unidad.
2. Actualiza y .
3. Continua el proceso hasta que se cumpla el criterio de parada.

# CAPITULO III. IMPLEMENTACIÓN DE LOS ALGORITMOS

### Mutador de Secuencias Simple

El Mutador de Secuencias Simple (MSS) desempeña un rol central en el proceso de manipulación de secuencias de ADN, proporciona un enfoque integral en la inserción aleatoria de gaps (mutaciones) al genoma de ADN del virus de VIH del simio, lo que permite a su vez generar poblaciones mutadas de esta secuencia, de manera eficiente.

A diagram of a simple process

Description automatically generated

Una de las ventajas clave del MSS es su capacidad para insertar y eliminar mutaciones de manera aleatoria a un genoma de ADN. Esto significa que no se limita a realizar cambios puntuales en la secuencia, sino que también le permite simular procesos naturales de mutación genética de una manera eficiente y versátil.

Esta característica es especialmente útil cuando se trabaja con poblaciones de secuencias y se busca explorar una amplia gama de posibles variantes genéticas.

Al permitir la inserción de gaps o espacios en posiciones seleccionadas al azar de una secuencia, el MSS brinda la flexibilidad necesaria para crear una diversidad de variantes genéticas. Estas variantes pueden ser cruciales para la solución del problema del AMS, además de la optimización y comparativa de los tres algoritmos previamente explicados y a su vez para las investigaciones relacionadas con la función de genes específicos, la respuesta a fármacos, la adaptación a entornos cambiantes y muchas otras áreas de estudio.

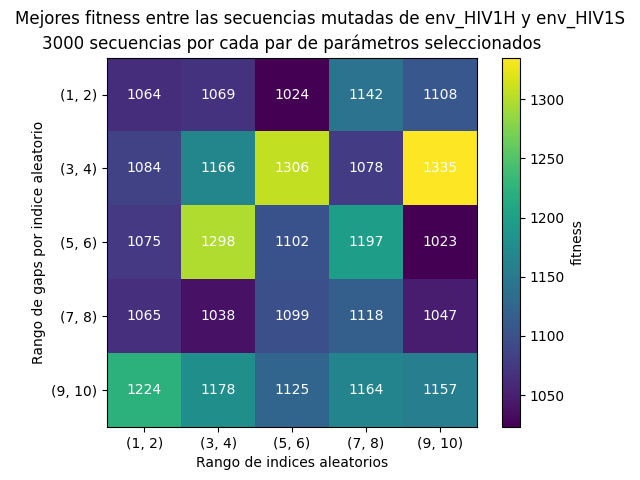
Es esencial no solo generar mutaciones en las secuencias de ADN, sino también garantizar que estas mutaciones sean relevantes y conduzcan a alineaciones significativas cuando se comparan con otras secuencias de referencia. Para lograr esto, es necesario ajustar cuidadosamente varios parámetros clave en el proceso de generación de mutaciones:

* **Número de Mutaciones:** En primer lugar, se debe determinar cuántas mutaciones se deben realizar en la secuencia original. Este número debe ser lo suficientemente alto como para introducir variabilidad genética, pero no tan alto que las secuencias mutadas sean completamente incoherentes con la secuencia de referencia.
* **Posiciones de Mutación:** Determinar las posiciones específicas dentro de la secuencia donde se introducirán las mutaciones. Esto puede hacerse de manera aleatoria o siguiendo un patrón específico. La elección de las posiciones de mutación puede influir en la similitud o diferencia con la secuencia de referencia.
* **Longitud de Gaps (Espacios):** Si se permiten gaps o espacios en la secuencia, es importante definir la longitud máxima de estos gaps. Esto afectará la extensión de las inserciones o deleciones permitidas.

Para determinar los valores óptimos de estos parámetros, se llevó a cabo un proceso experimental que involucró la generación de una población de secuencias mutadas. Este proceso se realizó utilizando las secuencias env\_HIV1H y env\_HIV1S, generando un total de secuencias mutadas, divididas de manera equitativa, es decir, secuencias mutadas para cada una de estas secuencias de referencia.

Durante este proceso, se variaron los parámetros seleccionados, y para cada conjunto de parámetros, se evaluó la calidad de las secuencias mutadas. El criterio de evaluación utilizado fue la aptitud o "fitness," que se definió en función del grado de alineamiento de los genes en las secuencias mutadas con respecto a las secuencias de referencia. En otras palabras, se buscó identificar las secuencias mutadas que lograron el mejor alineamiento con los genes de interés.

Este enfoque experimental permitió identificar los valores de parámetros que condujeron a las secuencias mutadas con el mejor fitness, lo que a su vez proporciona información valiosa para optimizar el proceso de generación de secuencias mutadas.



Los resultados derivados del análisis de 3000 secuencias generadas para cada conjunto de parámetros seleccionados revelan patrones significativos en cuanto a la obtención de alineaciones óptimas y subóptimas en las secuencias mutadas.

Estos resultados indican que, en la población de secuencias generadas, se suelen encontrar las mejores alineaciones de genes cuando se utilizan rangos de índices aleatorios entre y , así como rangos de gaps por índice aleatorio entre y .

## Algoritmo Genético

Se implementó el Algoritmo Genético basado en lo siguiente:

* Modificador Simple de Secuencias
* Modificador Cruce de Secuencias
* Proceso de Mutaciones

### Modificador Cruce de Secuencias

El modificador de cruce de secuencias (MCS) es un algoritmo genético basado en el problema de la mochila (knapsack problem). En este enfoque, se generan nuevas secuencias cruzando genes de secuencias previamente modificadas, utilizando un punto de cruce generado de manera aleatoria. Este algoritmo incorpora funcionalidades similares a las utilizadas en el problema de la mochila, su objetivo principal es aplicar estas técnicas al alineamiento de genes.

El MCS desempeña un papel complementario al modificador simple de secuencias (MSS) mientras que el MSS se encarga de generar una población de secuencias mutadas a partir de una secuencia original, el MCS trabaja con las secuencias previamente generadas por el MSS. En otras palabras, estos dos modificadores colaboran en el proceso de manipulación de secuencias de genes.

El proceso se inicia con el MSS, que crea una población de secuencias mutadas a partir de una secuencia de referencia específica. Luego, estas secuencias mutadas se utilizan como entrada para el MCS, este algoritmo cruza las secuencias previamente generadas, combinando segmentos de genes de diferentes secuencias mutadas de manera aleatoria, lo que resulta en una nueva población de secuencias mutadas con una mayor diversidad genética.

A diagram of a diagram

Description automatically generated

En conjunto, estos dos modificadores de secuencias trabajan de manera sinérgica para generar y diversificar las secuencias mutadas, lo que es esencial para la investigación en el campo del alineamiento de genes y la exploración de variantes genéticas relevantes, la combinación de estas estrategias permite abordar de manera efectiva la optimización y la diversificación de las secuencias en el contexto del alineamiento genético.

#### Secuencias generadas por el MSS

De manera ilustrativa, presentamos dos secuencias genéticas, cada una con su propia longitud y una disposición específica de genes. Estas secuencias actúan como punto de partida para un proceso de manipulación genética con el fin de generar variantes controladas:

* **Primera Secuencia:** -A-A-A-B-B-B-C-C-C-D-D-D-E-E-E-
* **Segunda Secuencia:** AAA-BBB-CCC-DDD-EEE

Es importante destacar que estas secuencias poseen diferencias tanto en su longitud como en la composición de genes presentes. Nuestro objetivo principal es crear variantes de estas secuencias sin alterar el orden de los genes existentes, pero permitiendo mutaciones (gaps) controladas mediante la introducción de un punto de cruce.

#### Proceso de Mutación implementado el MCS

Este proceso de mutación opera mediante la generación aleatoria de un punto de cruce, a continuación, se presentan las secuencias resultantes después de aplicar esta técnica con un punto de cruce generado al azar en el índice 6:

* **Primera Secuencia:** AAA-BBB-C-C-C-D-D-D-E-E-E
* **Segunda Secuencia:** -A-A-A-B-B-B-CCC-DDD-EEE

En el primer caso, observamos que la primera secuencia experimenta mutaciones a partir del punto de cruce. Antes del punto de cruce, los genes permanecen con las mutaciones implementadas por el MSS, y es después de este punto que algunos genes sufren nuevas mutaciones al implementar el MCS, creando una variante de la secuencia originalmente mutada.

En cuanto a la segunda secuencia, se presenta un fenómeno similar de mutación; sin embargo, los cambios se producen en diferentes ubicaciones debido al punto de cruce, nuevamente, antes del punto de cruce, la secuencia se mantiene sin modificaciones, y es después de este punto que se producen alteraciones en los genes.

Este proceso de modificación de secuencias es esencial para explorar y generar nuevas variantes genéticas mientras se preserva la estructura y el orden original de los genes.

## Algoritmo de Reacciones Químicas

### Modificador de Reacciones Químicas

El Modificador de Reacciones Químicas (MRQ) se basa en una fusión única de conceptos químicos y bioinformáticos y se ha adaptado específicamente para abordar el desafío crucial del alineamiento de secuencias genéticas. El enfoque se asemeja a una simulación de reacciones químicas, donde las secuencias de genes se comportan como moléculas. Cada secuencia se asocia con una energía potencial y cinética, lo que rige su estructura y dinámica a lo largo del proceso de alineamiento.

En este contexto, el modificador modela cuatro tipos fundamentales de interacciones:

* **Colisiones Ineficaces:** Simulan colisiones suaves que se asemejan a colisiones con una barrera energética. Cuando una secuencia de genes colisiona con una barrera, se permite un cambio gradual en su estructura, reflejando mutaciones leves en el genoma.
* **Descomposición:** Cuando una secuencia choca con una barrera de alta energía, se desencadena una reacción de descomposición. Esto resulta en la división de la secuencia en dos o más fragmentos, generando secuencias diferentes. La energía se redistribuye entre los fragmentos, manteniendo la energía total constante.
* **Colisiones Ineficaces entre Moléculas:** Similar a las colisiones ineficaces con una barrera, pero involucra la interacción de dos secuencias. Si ambas secuencias poseen suficiente energía, experimentarán un cambio en su estructura.
* **Síntesis:** En esta reacción, dos secuencias colisionan y se fusionan en una secuencia nueva. Esta reacción es altamente energética y genera una secuencia distinta a las originales.

El algoritmo evalúa la energía potencial entre las secuencias de genes y decide si permite estos cambios. La energía cinética juega un papel crítico, ya que las secuencias deben contar con suficiente energía para superar las barreras energéticas y cambiar su estructura.

En situaciones en las que la energía cinética podría resultar insuficiente, se emplea un "buffer de energía" para gestionar y posibilitar las modificaciones.

El proceso de modificación de reacciones químicas se implementa de manera conjunta con el Modificador Simple de Secuencias (MSS). El algoritmo inicia generando una población de secuencias con genes mutados utilizando el MSS. Luego, se seleccionan aleatoriamente dos secuencias de esta población con el propósito de llevar a cabo una reacción química. La elección de la reacción química específica se realiza de manera aleatoria entre las cuatro reacciones posibles. En caso de que la reacción química sea exitosa, se procede a actualizar la población de secuencias.

En cuanto a los valores predeterminados para el "buffer de energía" y el número de colisiones, se establecen en y , respectivamente. Estos valores se han consolidado como las recomendaciones óptimas que generalmente ofrecen los resultados más favorables.

## Algoritmo de Recocido Simulado

La idea fundamental detrás del Recocido Simulado es explorar el espacio de soluciones en busca de la mejor solución posible, evitando quedar atrapado en óptimos locales. Esto se logra mediante una búsqueda que acepta soluciones subóptimas con cierta probabilidad en las etapas iniciales y disminuye gradualmente esa probabilidad a medida que avanza el algoritmo. Esta característica permite al Recocido Simulado escapar de óptimos locales y buscar soluciones globales en el espacio de búsqueda.

El algoritmo comienza con una solución inicial y un parámetro de temperatura que controla la probabilidad de aceptar soluciones subóptimas. A medida que el algoritmo avanza, la temperatura se reduce gradualmente según un programa predefinido (por ejemplo, un enfriamiento exponencial).

A temperaturas más altas, es más probable que se acepten soluciones subóptimas, lo que permite explorar el espacio de búsqueda de manera más amplia. A temperaturas más bajas, la probabilidad de aceptar soluciones subóptimas disminuye, lo que hace que el algoritmo se enfoque en refinar la solución actual.

El proceso continúa hasta que se cumple un criterio de parada, que puede ser un número máximo de iteraciones, una temperatura mínima, o una mejora insuficiente en la solución. Al final de la ejecución, el Recocido Simulado devuelve la mejor solución encontrada durante todo el proceso.

### Recocido Simulado adaptado al Alineamiento de Secuencias

El algoritmo ha sido adaptado específicamente para el alineamiento de secuencias genéticas. Utiliza el enfoque de recocido simulado para generar soluciones alternativas a partir de una secuencia original, aplicando mutaciones controladas con el MS. El criterio de evaluación utilizado para determinar la calidad de las nuevas soluciones se basa en su alineación con una secuencia de referencia establecida inicialmente.

El proceso comienza con la secuencia original y aplica mutaciones utilizando el MS. Luego, se calcula el alineamiento de cada secuencia mutada con respecto a la secuencia de referencia utilizando un enfoque de alineación de genes.

Durante la ejecución del algoritmo, se lleva a cabo el proceso de recocido simulado, donde se aceptan o rechazan las nuevas soluciones en función de su capacidad para mejorar el alineamiento con la secuencia de referencia. Si una nueva solución presenta un alineamiento mejorado, se acepta automáticamente. En caso contrario, la aceptación se determina probabilísticamente, lo que permite al algoritmo explorar soluciones subóptimas en busca de una mejora global.

El proceso se repite hasta que se cumple un criterio de parada, como alcanzar una temperatura mínima o no lograr una mejora significativa en el alineamiento. El resultado final es la mejor secuencia encontrada durante la ejecución del algoritmo, la cual se considera como una solución óptima o subóptima en términos de alineación de genes.

# CAPITULO IV. EXPERIMENTACIÓN Y RESULTADOS

## Comparativa de Metaheurísticas

### Generación de Secuencias

Se construyó una lista de secuencias mutadas que varió en tamaño, comenzando con 200 secuencias y aumentando gradualmente en incrementos de 100 secuencias en cada iteración.

Se proyecta un total aproximado de 10,500 secuencias generadas por los modificadores que generan nuevas secuencias. No obstante, debido a la naturaleza del algoritmo de reacciones químicas, este proceso puede dar como resultado la generación o eliminación de secuencias. Como resultado final, se obtuvieron 10,179 secuencias.

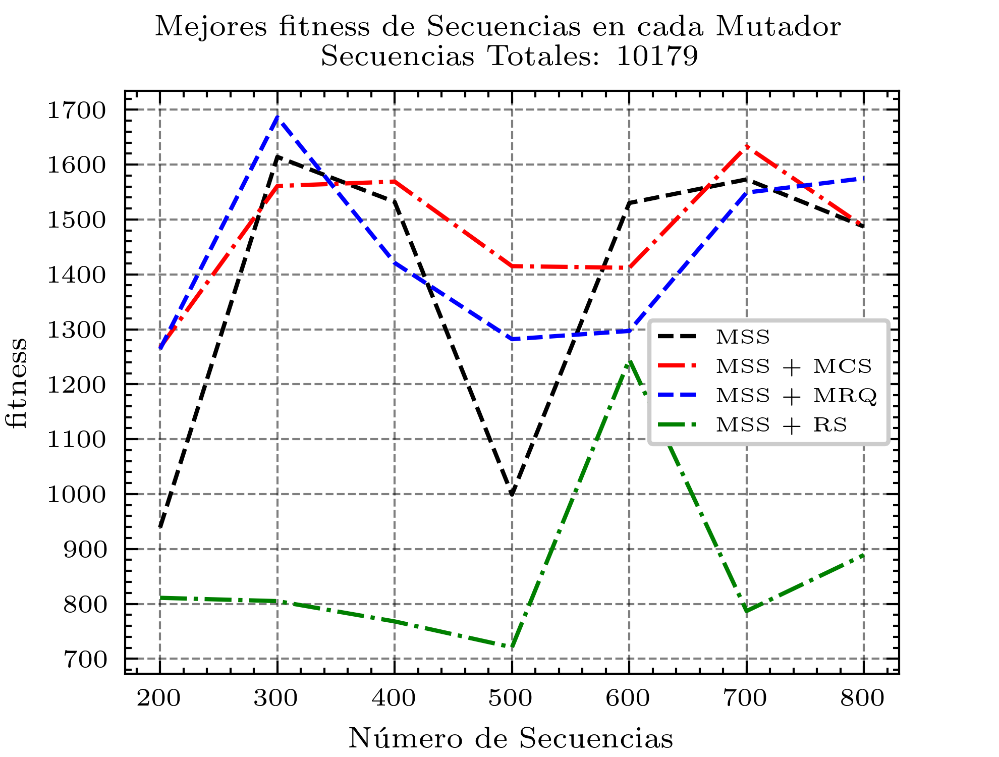
### Modificadores de Secuencias

El primer modificador de secuencias utilizado en este estudio fue el denominado "Modificador Simple de Secuencias" (MSS). Las secuencias generadas mediante este modificador se utilizaron como entrada para dos modificadores adicionales: el "Modificador Cruce de Secuencias" (MSC) y el "Modificador de Reacciones Químicas" (MRQ). Estos últimos generaron nuevas secuencias a partir de las generadas por el MSS.

Cada modificador de secuencias se implementó de manera independiente, y se registró el mejor valor de aptitud (fitness) obtenido para cada conjunto de secuencias generadas en cada iteración.

### Representación Gráfica y Numérica de Resultados Obtenidos

Los resultados de esta evaluación se presentan a través de una gráfica que ilustra la evolución de la aptitud en función del número de secuencias generadas. Asimismo, se proporcionan los resultados numéricos de interés.



### Resultados Numéricos

Tiempo Transcurrido (min): 87.34

* **Modificador Simple de Secuencias (MSS):** [939, 1614, 1532, 999, 1530, 1573, 1487]
* **MSS + Modificador Cruce de Secuencias (MSC):** [1267, 1561, 1569, 1415, 1412, 1633, 1487]
* **MSS + Modificador Reacciones Químicas (MRQ):** [1264, 1686, 1421, 1282, 1297, 1549, 1575]
* **MSS + Recocido Simulado (SR):** [811, 805, 768, 721, 1245, 787, 889]

### Media Aritmética y Desviación Típica de Aptitud

* **MSS:** 1382.0, 264.22
* **MSS + MSC:** 1477.71, 114.72
* **MSS + MRQ:** 1439.14, 154.63
* **MSS + SR:** 860.86, 163.71

### Análisis del Algoritmo de Recocido Simulado

En el marco de esta investigación, se llevó a cabo un riguroso análisis del algoritmo de Recocido Simulado, una de las metaheurísticas utilizadas en la optimización del alineamiento de genes en secuencias. Este algoritmo se caracteriza por su singularidad al emplear una única secuencia como base de operación, en contraste con otros modificadores que trabajan con múltiples secuencias. En esta sección, se presenta una evaluación detallada de los resultados obtenidos, así como una comparativa con otros modificadores de secuencias.

### Parámetros del Algoritmo

Para el funcionamiento del algoritmo de Recocido Simulado, se empleó una secuencia generada previamente por el "Modificador Simple de Secuencias (MSS)", la cual sirvió como punto de partida en cada iteración. Los parámetros utilizados para el algoritmo se seleccionaron cuidadosamente, basándose en experimentos previos que demostraron que ciertos valores tienden a proporcionar los mejores resultados en términos de la aptitud de los genes de las secuencias.

Los parámetros utilizados fueron los siguientes:

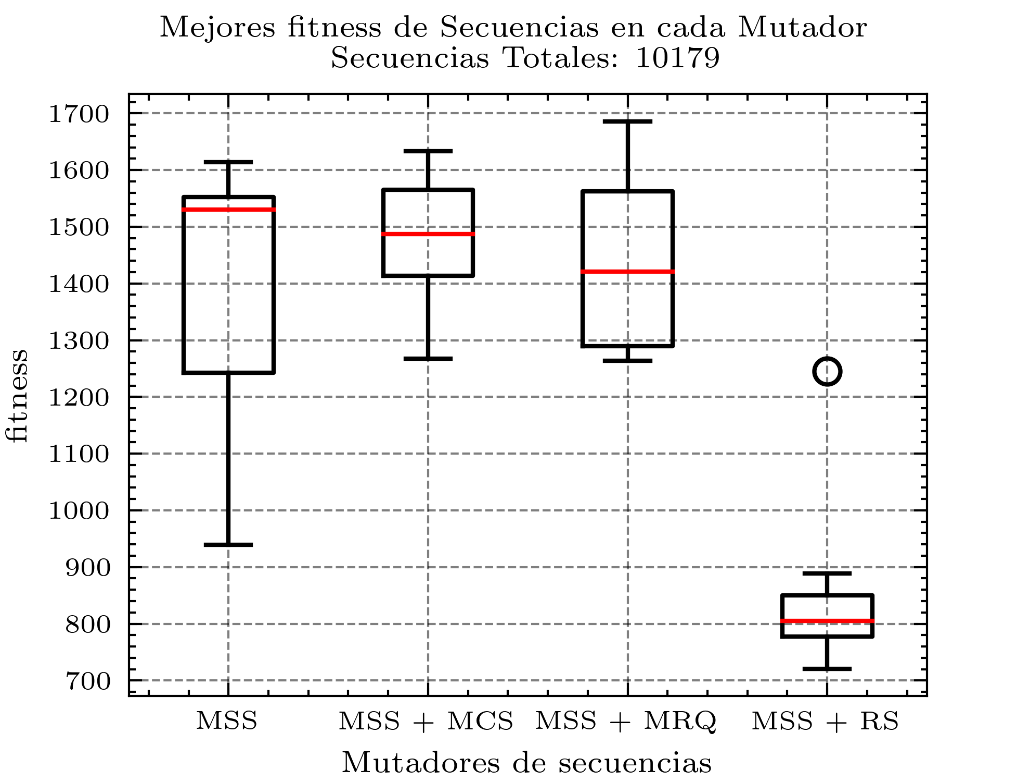
* **Temperatura Alta:** 1000
* **Temperatura Final:** 0.01
* **Tasa de Enfriamiento:** 0.99
* **Criterio de Metrópolis:** 100

Estos valores fueron elegidos tras un análisis exhaustivo y se mantuvieron constantes a lo largo de las iteraciones para garantizar la consistencia en la evaluación.

### Resultados y Comparativa

Los resultados obtenidos revelan que, en términos de mejor aptitud, el "Modificador de Reacciones Químicas" se destaca como el más efectivo. Sin embargo, en un análisis promedio, el "Modificador de Cruce de Secuencias" suele ofrecer resultados ligeramente superiores. Cabe destacar que el "Modificador de Recocido Simulado" trabaja principalmente con una sola secuencia, utilizándola como base para la generación de nuevas secuencias según su algoritmo interno.

Esta singularidad puede explicar por qué su rendimiento no alcanza el nivel de otros modificadores que operan con múltiples secuencias.



No obstante, es relevante señalar que el mejor valor de aptitud alcanzado por el algoritmo de Recocido Simulado se equipara al de otros modificadores de secuencias, sin embargo, los resultados del algoritmo no son tan satisfactorios en comparación con los demás modificadores, que, en la mayoría de los casos, superan sus resultados.

Este análisis y comparativa detallada aportan conocimientos valiosos en la selección de la metaheurística más adecuada para la optimización del alineamiento de genes en secuencias, contribuyendo al avance de la bioinformática y la genómica.

# REFERENCIAS BIBLIOGRAFICAS

* Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. Journal of Molecular Biology, 147(1), 195-197.
* Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 48(3), 443-453.
* Egea, J. A., & Riff, M. C. (2019). Bioinspired optimization methods for sequence alignment. Briefings in Functional Genomics, 18(6), 405-416.