UNIVERSIDAD AUTONOMA METROPOLITANA

Unidad Azcapotzalco

DIVISIÓN DE CIENCIAS BASICAS E INGENIERIA

MAESTRIA EN CIENCIAS DE LA COMPUTACIÓN

PROPUESTA DE PROYECTO

"Algoritmos Genéticos para el alineamiento de secuencias génicas en sistemática filogenética."

PRESENTA
L. C. I. Jorge Hernández Negrete

DIRECTOR
Dr. Katya Rodríguez Vázquez
Instituto de Investigación en Matemáticas Aplicadas y de Sistemas - UNAM
Dr. Javier Ramírez Rodríguez
Universidad Autónoma Metropolitana Azcapotzalco
Dr. Helga Ochoterena Booth
Instituto de Biología - UNAM

1. INTRODUCCIÓN.

Actualmente en el Área de Sistemática Filogenética existe el problema de que las herramientas usadas para realizar las alineaciones de secuencias génicas, no dan resultados eficientes. Lo anterior está relacionado con el hecho de que al elevar el número de secuencias por evaluar aumenta la complejidad del problema y consecuentemente aumenta el tiempo de procesamiento de manera exponencial, como se verá más adelante. Muchas veces, además, los resultados de las alineaciones son poco consistentes pues los programas que actualmente se utilizan dependen del orden de comparación entre las secuencias. Más aún, los algoritmos que actualmente se emplean para alinear secuencias pocas veces son consistentes en cuanto a los criterios que se utilizan posteriormente para el análisis filogenético de las secuencias. Por ejemplo, los algoritmos de alineación emplean criterios de costo arbitrario, mientras que los de análisis filogenético pueden emplear el criterio de parsimonia (reducción del número de explicaciones o eventos necesarios para postular una hipótesis) o el de probabilidad.

En este trabajo se hace una propuesta para realizar la alineación de secuencias génicas en forma eficiente, mediante la utilización de Algoritmos Genéticos y el Criterio de Eventos, el cual es un criterio propuesto recientemente [4]. En el contexto de alineamiento, se considera que un evento es un cambio dado por la transformación de una base a otra o por la pérdida/ganancia (inserción/delesión) de bases (que se representa como un “espacio” en la secuencia conocido como “gap”). Consideramos que los Algoritmos Genéticos tienen estructuras muy similares a las del problema en cuestión y sumado a su eficiencia probada, podremos encontrar mejores resultados comparados con los que se tienen actualmente.

Se definirá un procedimiento con el propósito de que mejore las alineaciones que se están obteniendo con las herramientas actuales y que realice búsquedas aleatorias inteligentes. La necesidad de implementar búsquedas aleatorias inteligentes tiene que ver con el hecho de que la evaluación de las posibles alineaciones entre un numero relativamente pequeño de secuencias requiere la comparación de un enorme numero de alineaciones posibles, lo que puede hacer que el problema no tenga solución exacta en tiempos finitos, como veremos mas adelante.

2. MARCO TEÓRICO, ANTECEDENTES Y REVISION BIBLIOGRAFICA.

2.1 Evolución Natural.

Charles Darwin y Wallace explicaron la evolución de las especies por selección natural en su estudio Teoría de la Evolución por Selección Natural. Sin embargo, se conocen otros estudios con ideas alternativas sobre los mecanismos involucrados en el proceso de cambio, como los de Demócrito (2600 a. e.) y Lamarck (1806) [11].

2.2 Algoritmos Genéticos.

A finales de los años 60 se realizaron investigaciones que pretendían explicar y modelar matemáticamente los procesos de la evolución natural. Esto dio origen a un campo nuevo llamado Computación Evolutiva [17]. Dentro de este campo están los Algoritmos Genéticos que se fundamentan en el estudio de la naturaleza y en el hecho de que se han formado organismos con una gran capacidad y complejidad que responden a su entorno de una manera muy eficiente. La naturaleza ya hizo las cosas y las hizo bien.

Los Algoritmos Genéticos fueron desarrollados por John Holland junto con sus alumnos y colegas en la Universidad de Michigan. Este tipo de algoritmos de búsqueda está basado en emulaciones de mecanismos de selección natural y genética natural. Combinan la sobrevivencia del más fuerte entre estructuras de cadenas con intercambios de información aleatoria. En cada generación, un conjunto nuevo de criaturas artificiales (cadenas) es creado usando bits y piezas del más fuerte de los anteriores. Los Algoritmos Genéticos explotan eficientemente información histórica para especular sobre puntos de búsqueda nuevos con un mejoramiento esperado [1].

2.3 Esquema del Algoritmo Genético

Los Algoritmos Genéticos son diferentes de los métodos normales de optimización y procedimientos de búsqueda [1].

1. Trabajan con una codificación del conjunto de parámetros, no con los parámetros mismos.
2. Buscan en una población de puntos, no en un punto simple.
3. Usan rentabilidad de información (función objetivo)
4. Usan reglas de transición probabilística, no reglas determinísticas.

La forma de funcionar de los Algoritmos Genéticos consiste en generar aleatoriamente una población inicial de individuos o soluciones, codificados de alguna manera, generalmente cadenas de unos y ceros, como se ve en la Figura 1.

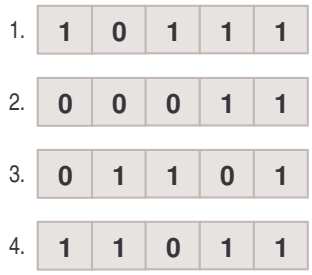


Figura 1.
Cada una de las cadenas de bits se llaman individuos y son las soluciones. Al conjunto se le llama población.

Posteriormente se crearán nuevas poblaciones usando 3 operadores genéticos que son una representación de la selección que existe en la naturaleza:

- Operador de selección
- Operador de cruza
- Operador de mutación

La selección escoge a los individuos de la población que son mejores con respecto a los otros mediante la evaluación de la función objetivo para que sean cruzados con otros para producir nuevos y mejores individuos.

La cruza intercambia partes de los individuos, así genera nuevos individuos, como se ve en la Fig. 2.

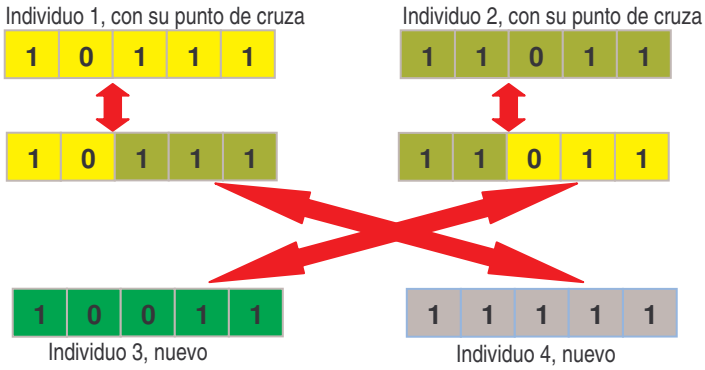


Figura 2. Proceso de cruza. Se selecciona un punto para cruzar y se hace el intercambio de la parte de ambos

La mutación cambia alguna parte o partes del individuo aleatoriamente, como se ve en la Figura 3.

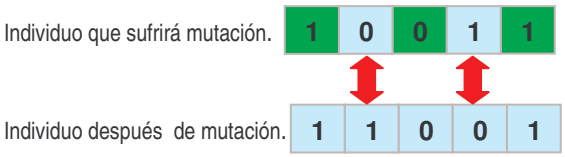


Figura 3. Proceso de mutación. Se selecciona los puntos donde se realizara la mutación.

2.4 Sistemática Filogenética.

En Biología hay una disciplina llamada Sistemática que estudia la clasificaron de los seres vivos y hay una ciencia llamada Filogenia que estudia las relaciones de género, especie, etc., entre las distintas especies, desde el origen de la vida en la tierra, hasta la actualidad [4,11]. Estos dos conceptos dan origen a la Sistemática Filogenética que se ocupa de proponer Hipótesis de Filogenia, las cuales se basan en la comparación de caracteres tanto morfológicos como de ADN, es decir, secuencias génicas (el ADN está compuesto por las llamadas bases nucleares, Adenina, Timina, Citosina y Guanina).

Un ejemplo de comparación de caracteres morfológicos es, si tenemos dos flores, una de color blanco y otra de color rojo, y ambas tienen la misma forma, se puede postular que tuvieron un ancestro común y en algún momento de la evolución sucedió un cambio. Ejemplo de comparación de ADN seria, si tenemos dos secuencias de ADN podemos determinar que tantas coincidencias o

diferencias tienen en cuanto a sus bases nucleares y de esta manera postular si están o no relacionadas filogenéticamente.

2.5 Alineamiento Genético

Para realizar la comparación por ADN se requiere obtener secuencias genéticas comparables propias de cada especie. Una vez que se tienen, éstas pueden ser de la misma longitud o pueden variar en longitud, como se ve en la figura 4. En la figura 5 podemos ver pequeño esquema de la sistemática filogenética.

```
ACTTCCGAATTTGGCT
ACTCGATTGCCT
```

Figura 4. Secuencias génicas comparables que varían en longitud

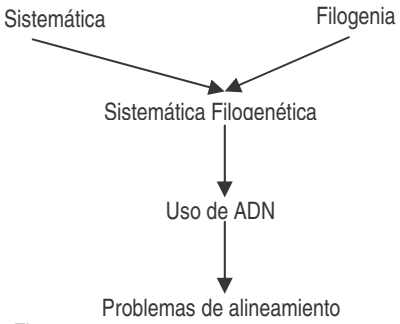


Figura 5.

Cuando esto sucede se presupone que, o una de las secuencias perdió un fragmento (delesión) o la otra ganó un fragmento (inserción), es decir, se debe presuponer que hubo un proceso que se resume como “indel”. Las indels se representan gráficamente con guiones que se conocen como “gaps”. Un gap es un espacio (es parte de un patrón), que refleja un fenómeno de indel (parte del proceso). La postulación de indels resulta en cadenas con patrones que tienen la misma longitud para las secuencias que inicialmente eran de diferente tamaño. A lo largo de este escrito utilizaremos el concepto de indel y gap de una forma equivalente. En la figura 6 se ve las dos secuencias de la figura 4 con tres posibles alineaciones, con lo cual las secuencias iniciales ya son de igual tamaño.

```
                        ACTTCCGAATTTGGCT
                        ACTCGATTGCCT

ACTTCCGAATTTGG-CT       ACTTCCGGAATTTGGCT       ACTTCCGGAATTTGGCT
|||  |||  |||  ||       |||*    *||| |*||       |||*     **|||*||
ACT--CGA--TTG-CCT       ACTC----GATT-GCCT       ACTC-----GATTGCCT
```

Fig. 6. Las secuencias génicas se alinean para encontrar sus patrones comparables.
En este ejemplo se postulan 3 posibles alineaciones.
Los “pipes” o líneas verticales, indican coincidencia.
Los asteriscos indican cambio de una base a otra.
Los guiones indican un indel.

A este proceso se le conoce como Alineamiento de Secuencias Génicas, dicho proceso consiste en hacer evaluaciones de dos o más cadenas para determinar qué es comparable entre ellas o qué tantas diferencias tienen, esto es, encontrar patrones en un conjunto de secuencias, y, posteriormente, identificar sus coincidencias ancestrales genéticas. En la figura 7, se puede ver una

alineación de secuencias génicas hecha a ojo por un experto y se puede notar un patrón a simple vista, por ejemplo, las Adeninas denotadas por la letra A y en color rojo aparecen en varias partes y debajo o encima de casi todas ellas hay otra adenina, lo cual también se ve con las demás bases. También se observa una región compartida por la mayoría de las secuencias que implica la inserción o delesión de seis bases nucleares, representando un gap. Es este gap el que permite las coincidencias a ambos lados de él. Por otro lado, se observan varias columnas para las cuales alguna o algunas secuencias no coinciden en cuanto a la base nuclear. Para estas secuencias es necesario postular un cambio de una base nuclear a otra.

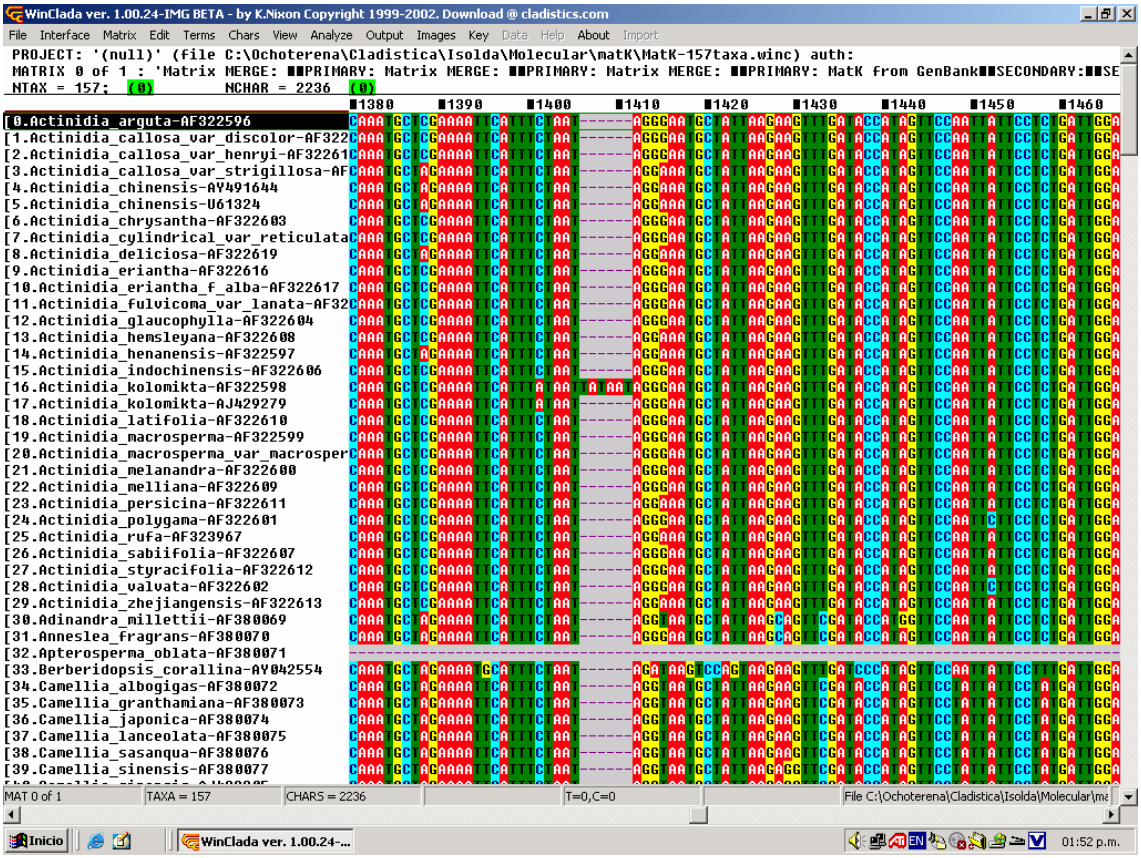


Figura 7. Ejemplo de alineación con un patrón visible.

El problema del alineamiento comienza por el hecho de que no existe una forma única de correlacionar a dos o más secuencias y en los casos relevantes no se puede conocer la historia evolutiva de las especies, solo se puede deducir a partir de la comparación de los patrones obtenidos con las observaciones. Por ejemplo, para el caso representado en la figura 6, se puede preguntar cuál de las alineaciones alternativas es la correcta o la mejor. Esta pregunta no la podemos responder, debido a que no sabemos realmente cómo fue la evolución, así que sólo podemos hacer deducciones mediante el establecimiento de criterios que nos llevarán a preferir alguna de las alineaciones en lugar de las otras [4].

2.6 Caso de estudio.

En el Instituto de Biología de la Universidad Nacional Autónoma de México, trabajan con secuencias génicas y consideran que las herramientas utilizadas para tal propósito, como Clustal (Clustal es un programa computación muy popular que realiza alineaciones por medio del criterio de costos), no

son suficientemente buenas, es decir, no es posible encontrar algún patrón lógico en el conjunto de secuencias alineadas y el tiempo necesario para realizar el procesamiento refleja que no se trata de una buena herramienta.

2.7 Criterio de Costos.

Las herramientas para alineación genética pueden usar varios tipos de criterios, el más usado es el criterio de costos, el cual incluye tres variantes principales:

a) Costos para minimizar sustituciones (este da un costo mayor a las sustituciones para que haya menos)
b) Costos para minimizar indels (este le da un costo mayor a los indels para que haya menos)
c) Costos para balancear sustituciones e indels (este da un costo equivalente a sustituciones e indels esperando tener un numero similar de ambos).

Estos costos son asignados subjetivamente, lo cual resulta en alineaciones que pueden variar mucho, dependiendo de quien las haga. Otra situación problemática consiste en usar el tipo de costo del inciso c), que balancea las sustituciones e indels, con el cual, lógicamente, puede resultar en varias alineaciones con el mismo valor, lo cual complica la interpretación de la información filogenética útil. Por ejemplo en la figura 8 se tienen dos secuencias y en la figura 9 se postulan dos posibles alineaciones que se evaluarán usando el criterio de costos del inciso c), donde el costo para una sustitución es igual al de un gap, es decir igual a 1, y el costo por la extensión de un gap o gap contiguo, es 0, esto dará un resultado igual lo cual dificulta elegir alguna.

```
ACTTCCGAATTTGGCT
ACTCGATTGCCT
```

Figura 8.

Alineación	Costo Sustitución = 1	Costo inicio Gap = 1	C. extensión Gap = 0	Costo final
ACTTCCGAATTTG-GCT ||| ||| || | || ACT--CGA-TT-GC-CT	1	1	0	0(1)+5(1)+1(0)=5
ACTTCCGAATTTGGCT |||* *||| |*|| ACTC---GATT-GCCT	1	1	0	3(1)+2(1)+2(0)=5

Figura 9. Se da un costo igual a una sustitución y a un gap, y obtendremos un costo final igual.

En la figura 10 se verá que se variaron los costos de la figura 9, de acuerdo con el criterio del inciso b), que mantendrá más sustituciones que gaps, para lograr una diferencia que nos permita preferir una de las dos alineaciones. Ahora hay elementos para preferir la segunda alineación porque tiene menor costo.

Alineación	Costo Sustitución = 1	Costo inicio Gap = 2	C. extensión Gap = 0	Costo final
ACTTCCGAATTTG-GCT ||| ||| || | || ACT--CGA-TT-GC-CT	1	2	0	0(1)+5(2)+1(0)=10
ACTTCCGAATTTGGCT |||* *||| |*|| ACTC---GATT-GCCT	1	2	0	3(1)+2(2)+2(0)=7

Figura 10. Se da un costo menor a una sustitución que a un gap o indel.

El criterio del inciso b) es mas usado por los biólogos, que prefieren poner el costo de un gap mayor al de una sustitución, pues de lo contrario el alineamiento podría resultar en comparaciones triviales, donde todo se podría explicar con un gran número de indels.

Sin embargo, en la figura 11 podemos apreciar que al utilizar una tercer alineación con un criterio que minimiza indels (gaps), se preferirá éste resultado por tener menor costo, esto quiere decir que el orden en que se meten las secuencias repercute en el resultado.

Alineación	Costo Sustitución = 1	Costo inicio Gap = 2	C. extensión Gap = 0	Costo final
ACTTCCGAATTTG-GCT ||| ||| || | || ACT--CGA-TT-GC-CT	1	2	0	0(1)+5(2)+1(0)=10
ACTTCCGAATTTGGCT |||* *||| |*|| ACTC---GATT-GCCT	1	2	0	3(1)+2(2)+2(0)=7
ACTTCCGAATTTGGCT |||* **||*|| ACTC----GATTGCCT	1	2	0	4(1)+1(2)+3(0)=6

Figura 11. Se da un costo igual a una sustitución y a un gap o indel}

La comparación por pares es un problema de los algoritmos actuales, pues si hay errores en las comparaciones intermedias, estos se acarrean a las siguientes, lo que a menudo resulta en patrones como los que se ven en la figura 12, donde claramente hay dos secciones discordantes entre sí.

Figura 12.

O como lo que se ve en la figura 13, donde no se ve un patrón claro.

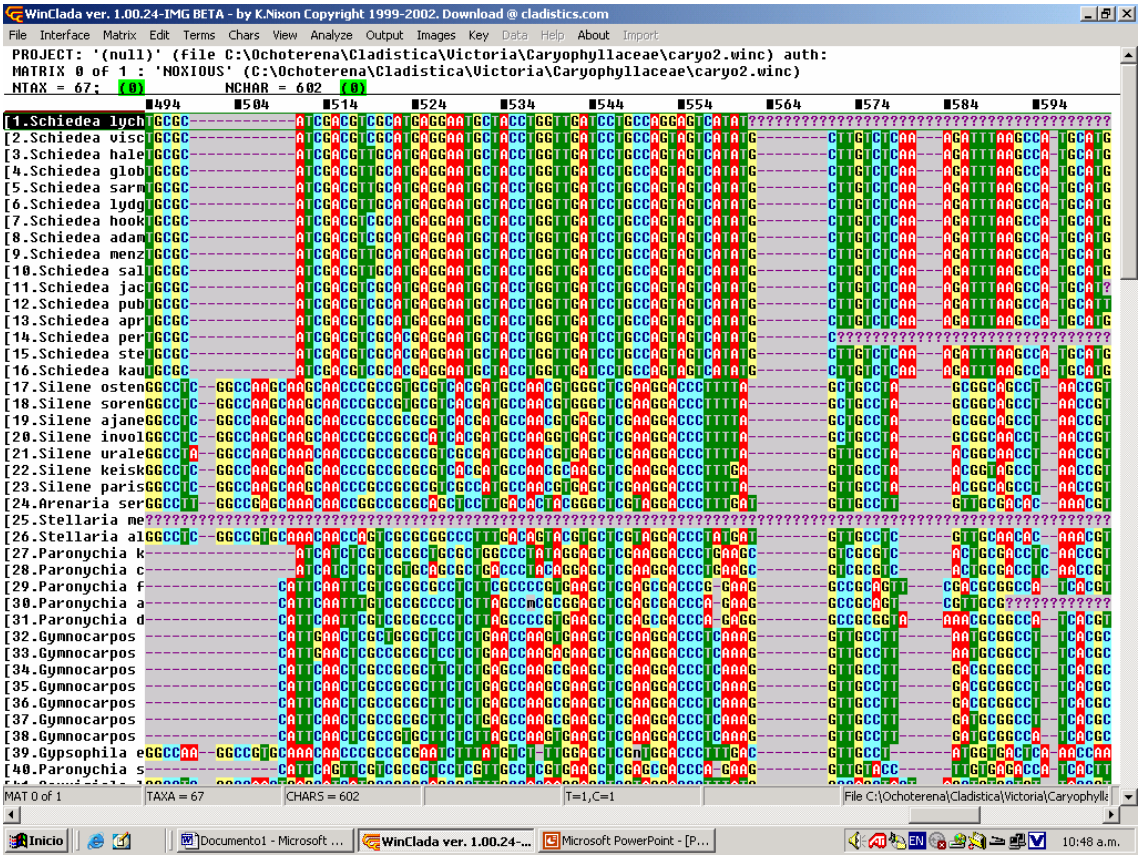


Figura 13.

La principal desventaja de este enfoque es que se puede quedar atrapado en algún mínimo local, esto lo podemos ver en la figura 13, con las varias secciones dispares que se ven. Esto proviene de la naturaleza glotona del algoritmo (se toma el primer elemento y se compara contra todos los demás elementos, se toma el segundo elemento y se compara contra todos los demás elementos, y así sucesivamente [8]). Esto significa que si algún error es hecho en alguna de las alineaciones intermedias, este no puede ser corregido después que se suman más secuencias [7], pues no se realiza un reajuste global, como es el caso de una herramienta muy popular llamada Clustal.

Existen otros criterios para alinear, el criterio de similitudes y el de modelos, que básicamente trabajan igual que el de costos o tienen otros problemas pero no entran en nuestro caso de estudio.

3. JUSTIFICACIÓN

3.1 El Tamaño del Problema.

Sean las dos secuencias de la figura 14. Dichas secuencias se pueden visualizar como una matriz de dos dimensiones porque hay que comparar cada posición en una de las secuencias contra cada posición en la otra, como se ve en la figura 15

```
ACCGGTCAA
ACCGT
```

Figura 14.

	A	C	C	G	G	T	C	A	A
A									
C									
C									
G									
T									

Figura 15.

En la figura 15, es obvio que la primera posición de ambas secuencias coincide, así que no necesitamos usar un gap o mutación para explicarla. Es solo hasta la quinta posición que tenemos un desempate, es decir, las bases son diferentes, como se ve en la figura 16, entonces se tienen dos opciones:

a) Se puede preferir postular una mutación, como se ve en la figura 16.

	A	C	C	G	G	T	C	A	A
A	A								
C		C							
C			C						
G				G					
T					T				

Figura 16.

b) Se puede preferir postular un indel, es decir, incorporar un gap [4] [8]. En la figura 17 se ve que la secuencia en el eje vertical se ha desplazado hacia abajo.

	A	C	C	G	G	T	C	A	A
A	A								
C		C							
C			C						
G				G					
-					-				
T						T			

Figura 17.

El problema se complica porque:

1) No se deben eliminar datos de ninguna de las secuencias, es decir, no se deben borrar bases.
2) La secuencia mas larga también puede sufrir indels (gaps) [4].

Los dos puntos anteriores se ejemplificaran tomando las dos secuencias de la figura 17 las cuales tienen el mismo número de bases. Y en la figura 18 se muestra como ninguna perdió alguna de sus bases y como se esta insertando un gap en ambas.

```
ACCGGTCTAA
ACCGTCTAGA
```

Figura 17.

	A	C	C	G	G	T	C	T	A	→	A
A	A										
C		C									
C			C								
G				G							
↓					-						
T						T					
C							C				
T								T			
A									A		
G										-	
A											A

Figura 18.

En casos reales se deben de probar, tantas combinaciones de desplazamientos como sea posible, se tienen secuencias largas (de hasta miles de posiciones) y se tienen que comparar más de dos secuencias. En el caso con más de tres secuencias (los casos reales) se tendrían que usar matrices multidimencionales, lo cual seria imposible [4] [8]. El caso se complica aun más por la existencia de más de una solución posible, como se ve en la figura 15 con sus dos posibilidades, y por la posibilidad de gaps contiguos, es decir, aquellos que ocupan más de una posición seguida [4].

3.2 Espacio de Posibles Soluciones.

En este punto cabria preguntar que tamaño tiene el espacio de posibles soluciones del cual hablamos.

Supongamos que tenemos dos cadenas como las que se ven en la figura 19, una más grande que la otra, la más grande tiene dos bases y la más corta una. También supongamos que la cadena más grande solo puede crecer en gaps un número igual al de la cadena más chica, entonces en este caso solo crecería uno.

```
GT
T
```

Figura 19.

Entonces, para obtener todas las posibles alineaciones de las cadenas de la figura 19, tendríamos que sacar la multiplicación de las combinaciones de:

$(n + k)C_n$, para la cadena más grande y de:

$(n + k)C_m$, para la cadena más corta [21].

Donde: n = al número de bases de la cadena mas larga
m = al número de bases de la cadena mas corta
k = al número de elementos que se pueden agregar a la cadena mas grande.

Si substituimos valores con respecto a las cadenas de la figura 19 se tendrá que:

n = 2
m = 1
k = m = 1

$$\binom{(n+k)}{n} \times \binom{(n+k)}{m} = \binom{(2+1)}{2} \times \binom{(2+1)}{1}$$
$$= \binom{3}{2} \times \binom{3}{1}$$
$$= 3 \times 3 = 9$$

Entonces el número total de soluciones para las cadenas de la figura 19 es igual a 9, y se muestran en la figura 20.

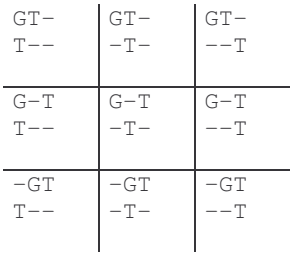


Figura 20.

Ahora supongamos que las secuencias de la figura 19 aumentan una base cada una, como se ve en la figura 21.

Figura 21.

Si substituimos valores con respecto a las cadenas de la figura 21 se tendrá que:

n = 3
m = 2
k = m = 2

$$\binom{(n+k)}{n} \times \binom{(n+k)}{m} = \binom{(3+2)}{3} \times \binom{(3+2)}{2}$$
$$= \binom{5}{3} \times \binom{5}{2}$$
$$= 10 \times 10 = 100$$

Entonces el número total de soluciones para las cadenas de la figura 21 es 100.

Ahora supongamos que utilizamos una cadena de tamaño real de 600 bases (Las cadenas reales podrían medir alrededor de 500 a 2000 bases), y esta puede aumentar un 10 %. Para calcular tan solo las combinaciones de esta cadena, habría que manejar un espacio de posibles soluciones de:

$$\binom{(n+k)}{n} = \binom{(600+60)}{600}$$
$$= \binom{660}{600} = 1.124849414603962 \times 10\ E86$$

Si suponemos que una computadora realiza 10 E9 (léase 10 exponente 9) operaciones por segundo, se tendría que el proceso tardaría 3.566874093 x 10 E69 años [21].

Con los sencillos ejemplos anteriores se pretende mostrar por qué el problema de alineamiento en casos más complejos (casos reales) no tiene una solución exacta en tiempos finitos, no seria factible poder analizar todas las alineaciones posibles. Por lo anterior, se obliga a buscar métodos alternativos, en este caso utilizaremos Algoritmos Genéticos, pues éstos utilizarían una estrategia inteligente de búsqueda que no requeriría de la evaluación de todas las posibles combinaciones de alineamientos al descartar conjuntos de alineamientos poco óptimos, para poder resolver este problema en tiempo razonable y de una forma eficiente.

3.3 Criterio de Eventos

El Criterio de Evento se ilustra de la siguiente manera [4], un evento es un cambio dado por la transformación de una base a otra, la cual gráficamente la podemos ver con un asterisco en la figura 22:

```
ACTT
|||*
ACTG
```

Figura 22.

Un evento también puede estar dado por la pérdida/ganancia de bases (que se representa como un gap, el cual puede implicar una sola base o más de una base), esto lo podemos ver gráficamente con los guiones en la figura 23, donde los guiones seguidos cuentan como un solo evento.

```
ACTTCCGAATT
|||   ||| ||
ACT--CGA-TT
```

Figura 23.

Con lo anterior, la Función de Eventos minimizaría el número de sucesos necesarios para explicar la coincidencia entre dos secuencias. Debido a que este caso puede resultar en un alineamiento trivial, para el cual se considere una secuencia de manera completamente adyacente a la otra, es necesario introducir como parte del criterio de eventos el número de posiciones variables, es decir, el objetivo es minimizar la combinación de cambios (gaps o cambios de una base nuclear a otra) y número de posiciones variables de la siguiente manera:

Min (FE = E (S, I) + NPV)

Donde: FE es la Función de Eventos que será minimizada.
E(S, I) es el número de eventos dado por sustituciones e indels.
NPV es el número de posiciones variables, es decir, donde hay cambios.
S es el número de sustituciones.
I es el número de indels

Ahora con las secuencias de la figura 24, se verá como se utiliza el criterio de costos y el criterio de eventos.

```
GTTCC
TTC
```

Figura 24.

Haciendo la alineación por costos para las secuencias de la figura 24, arbitrariamente podemos suponer que:

- Los cambios de una base a otra valen 5.
- La apertura de un gap cuesta 10.
- La continuación de un gap cuesta 3.

En la figura 25 se ve los resultados obtenidos para tres posibles alineaciones utilizando criterio de costos, también se ve que se escogería la alineación del inciso c), por tener menor costo, y las demás alineaciones se trabajarían sobre ésta, dando los resultados ya mencionados.

La figura 25 también nos muestra el resultado obtenido utilizando el criterio de eventos, con el cual se escogería la alineación del inciso a).

a)	b)	c)

12345
GTTCC
  |||
-TTC-
	GTTCC
	TTC
	
	12345
	GTTCC
	|
	TTC--
		12345
		GTTCC
		  |*|
		--TTC

Criterio Costos normal: 0(5)+2(10)+0(3)= 20	Criterio Costos normal: 2(5)+1(10)+1(3)= 23	Criterio Costos normal : 1(5)+1(10)+1(3)= 18
Criterio Eventos + posiciones variables: 2+2= 4	Criterio Eventos + posiciones variables:3+4= 7	Criterio Eventos + posiciones variables: 2+3= 5

Figura 25.

a) Dos eventos en la posición 1 y 5, mas dos posiciones variables en la posición 1 y 5, y daría cuatro.
b) Tendría tres eventos en la posición 1, 3 y de 4 a 5 (los gaps que hay en las posiciones 4 y 5 como son continuos se tomaran como un evento), mas cuatro posiciones variable en las posiciones 1, 3, 4, 5, esto daría siete.
c) Tendría dos eventos en las posiciones 1 a 2 y en la 4 (de igual forma como los gaps de las posiciones 1 y 2 son continuos se toman como uno), mas tres posiciones variables que estarían en las posiciones 1, 2 y 4, esto daría cinco.

Ahora, en la figura 26, se agregará una tercera cadena a las cadenas de la figura 24 con dos posibilidades, lo cual aumentara a 6 el número de posibles alineaciones, y se hará el mismo proceso de la figura 25, así se podrá comparar los resultados obtenidos tanto por criterio de costos como por criterio de eventos. Cabe destacar que se verá que el número de eventos puede variar.

a)	b)	c)

12345
GTTCC
  |||
-TTC-
  ||||
GTTC-
	GTTCC
	TTC
	GTTC
	12345
	GTTCC
	|
	TTC--
	| |
	GTTC-
		12345
		GTTCC
		  |*|
		--TTC
		  |*
		GTTC-

Criterio Eventos + posiciones variables: 2+2= 4	Criterio Eventos + posiciones variables: 4+4= 8	Criterio Eventos + posiciones variables: 3+4= 7

Figura 26.

Continuación Figura 26.

d)	e)	f)

12345
GTTCC
  |||
-TTC-
  |
-GTTC
	12345
	GTTCC
	|
	TTC--
	***
	-GTTC
		12345
		GTTCC
		  |*|
		--TTC
		  |||
		-GTTC

Criterio Eventos + posiciones variables: 4+4= 8	Criterio Eventos + posiciones variables: 6+5= 11	Criterio Eventos + posiciones variables: 4+3= 7

a) Dos eventos en la posición 1 y 5 (los gaps verticales se tomarían como un evento), y dos posiciones variables en la posición 1, 5, lo que daría 4.
b) Cuatro eventos en las posiciones 1, 3, de 4 a 5 y 5 (los gaps no coinciden en el mimo principio y fin, por lo que se reconocen como dos eventos) y 4 posiciones variables, lo que daría 8.
c) Tres eventos en la posición 1 a 2, 4 y 5, y 4 posiciones variables en 1, 2, 4 y 5, lo que da 7.
d) Cuatro eventos en la posición 1, 2, 3 y 4, y 4 posiciones variables en 1, 2, 4 y 5, lo que da 8.
e) Seis eventos en la posición 1 (dos eventos), 2, 3, 4 y 4 a 5, y 5 posiciones variables, lo que daría 11.
f) Tres eventos en las posiciones 1, 1-2, 2, y 4, y 3 posiciones variables, lo que daría 7.

Si se agregara una cuarta secuencia, no se compararían con los alineamientos de los incisos b), d) ni e), pues estos resultaron con los mayores valores para la función Min (FE = E (S, I) + NPV). Se tendría que comparar las posibles combinaciones cuando mucho sólo con los incisos a), c) y f), entonces los eventos quedarían como se ve en la figura 27:

GTTCC
TTC
GTTC
GTC

a)	a')	a'')	a''')

12345
GTTCC
  |||
-TTC-
  |||
GTTC-
|| |
GT-C-
	12345
	GTTCC
	  |||
	-TTC-
	  |||
	GTTC-
	| |||
	G-TC-
		12345
		GTTCC
		  |||
		-TTC-
		  |||
		GTTC-
		  **
		--GTC
			12345
			GTTCC
			  |||
			-TTC-
			  |||
			GTTC-
			||* |
			GTC--

Eventos + posiciones variables: 3+3= 6	Eventos + posiciones variables: 3+3= 6	Eventos + posiciones variables: 5+5= 10	Eventos + posiciones variables: 4+4= 8
c)	c')	c'')	c''')

12345
GTTCC
  |*|
--TTC
  |*
GTTC-
|| ||
GT-C-
	12345
	GTTCC
	  |*|
	--TTC
	  |*
	GTTC-
	| |||
	G-TC-
		12345
		GTTCC
		  |*|
		--TTC
		  |*
		GTTC-
		  **
		--GTC
			12345
			GTTCC
			  |*|
			--TTC
			  |*
			GTTC-
			||* |
			GTC--

Eventos + posiciones variables: 4+5= 9	Eventos + posiciones variables: 4+4= 8	Eventos + posiciones variables: 4+5= 9	Eventos + posiciones variables: 5+5= 10

Figura 27.

Continuación Figura 27.

f)	f')	f'')	f''')

12345
GTTCC
  |*|
--TTC
  |||
-GTTC
 * *
GT-C-
	12345
	GTTCC
	  |*|
	--TTC
	  |||
	-GTTC
	  |*
	G-TC-
		12345
		GTTCC
		  |*|
		--TTC
		  |||
		-GTTC
		  *||
		--GTC
			12345
			GTTCC
			  |*|
			--TTC
			  |||
			-GTTC
			 **
			GTC--

Eventos + posiciones variables: 6+5= 11	Eventos + posiciones variables: 5+4= 9	Eventos + posiciones variables: 5+4= 9	Eventos + posiciones variables: 6+5= 11

a) Tres eventos en las posiciones 1, 3 y 5, y 3 posiciones variables, lo que daría 6.
a') Tres eventos en las posiciones 1, 2 y 5 y 3 posiciones variables, lo que daría 6.
a'') Cinco eventos en las posiciones 1, 1-2, 3, 4 y 5, y 5 posiciones variables, lo que daría 10.
a''')Cuatro eventos en las posiciones 1, 3, 4-5 y 5, y 4 posiciones variables, lo que daría 8.

c) Cuatro eventos en las posiciones 1-2, 3, 4 y 5, y 5 posiciones variables, lo que daría 9.
c') Cuatro eventos en las posiciones 1-2, 2, 3 y 4 y 4 posiciones variables, lo que daría 8.
c'') Cuatro eventos en las posiciones 1-2, 3, 4 y 5 y 5 posiciones variables, lo que daría 9.
c''') Cinco eventos en las posiciones 1-2, 3, 4, 4-5 y 5 y 5 posiciones variables, lo que daría 10.

f) Seis eventos en las posiciones 1, 1-2, 2, 3, 4 y 5, y 5 posiciones variables, lo que daría 11.
f') Cinco eventos en las posiciones 1-2, 1, 2, 4 y 5, y 4 posiciones variables, lo que daría 9.
f'')Cinco eventos en las posiciones 1-2, 1, 2, 3 y 4, y 4 posiciones variables, lo que daría 9.
f''') Seis eventos en las posiciones 1-2, 1, 2, 3, 4 y 4-5, y 5 posiciones variables, lo que daría 11

Es importante notar que en cada cadena se obtiene un alineamiento que minimiza el número de eventos de manera global, no solo para un par de secuencias o grupos de secuencias comparadas, por lo que a simple vista se ve que se está generando un patrón dentro del subconjunto de alineamientos, como se puede ver en el inciso a) de la figura 27.

También en la figura 27 se observa que los alineamientos del inciso a) y a') tienen el mismo valor y tienen el menor valor con respecto a los demás, este caso implicaría la presentación de ambos alineamientos como resultado final del análisis. Esto representaría también una diferencia importante con respecto a los algoritmos que actualmente se usan para alinear secuencias pues, debido a que hacen comparaciones por pares y a que son dependientes del orden de adición de las secuencias, siempre arrojan un solo alineamiento, aunque existan otras posibilidades igualmente costosas, lo que resulta en que el mismo conjunto de secuencias pueda retribuir alineamientos distintos aun cuando se hayan utilizado exactamente los mismos parámetros.

3.4 Algunos Elementos para el Diseño del Algoritmo Genético.

El Algoritmo Genético estaría construido de la siguiente manera.

Las entradas o cadenas iniciales tendrán longitud variable.

La Función Objetivo estar dada por

$$\text{Min } (FE = E (S, I) + NPV)$$

La codificación de las soluciones estaría dada por la siguiente tabla:

A – 1
T – 2
C – 3
G – 4
(Gap) – 5

Entonces una cadena como esta: 11553442, representaría: AA - - CGGT

El operador de cruza estaría dado en un punto donde inicien y terminen con el mismo número de elementos en ambos segmentos, por ejemplo en las dos cadenas siguientes el punto de cruza estaría en los dos últimos números:

11553442
53133542

115534 I 42
531335 I 42

El operador de mutación estaría dado por el operador de inserción o eliminación de gaps, por ejemplo si se tienen las cadenas siguientes de diferente tamaño, la mutación estaría dada por la inserción o eliminación de gap.

11224321
4321

11224321
55432155

3.5 Propuesta.

La propuesta es pues, atacar los problemas de alineamiento de secuencias génicas, implementando un programa de cómputo que incorpore Algoritmos Genéticos y que cambie la función de costos por una función de eventos [4].

Este Criterio por Eventos es consistente con el siguiente paso que es la reconstrucción filogenética [4], la cual no tocaremos, y consistiría en la implementación de un criterio que ya se está empleando de manera manual por algunos expertos en el tema.

Se considera que los Algoritmos Genéticos podrán encontrar de una manera más eficiente y natural las alineaciones buscadas, presentando alineaciones mejores pues el algoritmo tomará y evaluará en conjunto las secuencias. También reducirán los tiempos por su paralelismo implícito [1,4,5,6,7,10]. Cabe mencionar que el estudio podría ser para cualquier área que tenga secuencias genéticas que quieran ser alineadas y que de lograrse este proyecto se estaría introduciendo una nueva forma de realizar las alineaciones en la Sistemática Filogenética.

Por último se muestra un ejemplo de tres secuencias que serán alineadas a manera de prueba inicial, las cuales corresponden a secuencias reales y están compuestas de 627-799 bases nucleares.

>Ardisia_crenata-AF547796
TGGAAACCTACTAAGTGAGAACTTTCAAATTCAGAGAAACCCTGGAATTAATAAAAATGGGCAATCCTGAGCCAAATCCTC
TTTTTCGAAAACAAAGATTAAAGGAAAATAAAAAAGAGGGATAGGTGCAGAGACTCAATGGAAGCTGTTCTAACAAATGGA
GTTGATCGCGTTGGTAGAGGAATCTTTTCATCAAAACTTCAGAAAGGATGAAAGATAAACGTATATACATATGCATATGTA
CTGAAATCCTATATCAAAATAAAATGCTTATTTTTTCTTTCTATGAAAAATAGAAGAATCGTTGCGAATCGATTCCACATT
CCACATTGAATAAAAAATTTCATATTCATTGATCAAATCATTTACTCCATAGTCTAATAGATCTTGTGAATAACTGATTAA
TCAGGCAAGAATAAAGATAGAGTCCCATTCTACATGTCAATACCGACAACAATAATGAAATTTATAGTACGAGGAAAATCC
GTCGACTTTATAAATCGTGAGGGTTCAAGTCCCTCTATCCCCAAAAGTTTATTTGACTCCCTAACTATTTATCCTATGCTA
TTTATACTATTACTATACTCCTTTTCGTTATTTGAGCAAGGAATCCCCGTTTGAATAATTCACGGTCCATATCATTATTCG
TACTGAAACTTACACAATTTTCACAATTTTCCTTTTTTTGAAAATCCAAGAAATCGCAGGGCCCACATAAGACTTTAATAA
TACTTTTTGTTTTTTAATTGACATAGACCCAAGTCATCTAGTAAAATGAGGATGATGCATCGGGGGTGGT

>Ardisia_gigantifolia-AF547795
TGGAACCTACTAAGTGAGAACTTTCAAATTCAGAGAAACCCTGGAATTAATAAAAATGGGCAATCCTGAGCCAAATCCTCT
TTTTCGAAAACAAAGATTAAAGGAAAATAAAAAAGAGGGATAGGTGCAGAGACTCAATGGAAGCTGTTCTAACAAATGGAG
TTGATCGCGTTGGTAGAGGAATCTTTTCATCAAAACTTCAGAAAGGATGAAAGATAAACGTATATACATATGCATATGTAC
TGAAATCCTATATCAAAATAAAATGCTTATTTTTTCTTTCTATGAAAAATAGAAGAATCGTTTCGAATCGATTCCACATTC
CACATTGAATAAAAAATTTCATATTCATTGATCAAATCATTTACTCCATAGTCTAATAGATCTTGTGAATAACTGATTAAT
CAGGCAAGAATAAAGATAGAGTCCCATTCTACATGTCAATACCGACAACAATAATGAAATTTATAGTACGAGGAAAATCCG
TCGACTTTATAAATCGTGAGGGTTCAAGTCCCTCTATCCCCAAAAGTTTATTTGGCTCCCTAACTATTTATCCTATGCTAT
TTATACTATTACTATACTCCTTTTCGTTATTTGAGCAAGGAATCCCCATTTGAATAATTCACGGCCCATATCATTATTCGT
ACTGAAACTTACACAATTTTCCTTTTTTTGAAAATTCAAGAAATCGCAGGGCCCACATAAGACTTTAATAATACTTTTTGT
TTTTTAATTGACATAGACCCAAGTCATCTAGTAAAATGAGGATGATGCATCGGGAGTGGT

>Ardisia_crenata-AF547730
TCAAAACCTGCATAGCAGAACGACCCGTGAACTTGTCTATACATAGGGGACGCATCGGATGGGTCCTTGACCCTCCGGTGT
CATCCCCTGCTAGTGGGTGAGCTCGTTCTCCGTCTCGGTTGCGAGTTCACTTTCCTAGTCGAACAACGAACCCCGGCGCGA
ACTGCGCCAAGGAAATCAAACAAAGAGATCGCACCCCTTACTCCTGTGTGCGGGTTGTGAGGGGTGATAGAATCTTGCATA
TAACAAAACGACTCTCGGCAACGGATATCTAGGCTCTCGCATCGATGAAGAACGTAGCAAAATGCGATACTTGGTGTGAAT
TGCAGAATCCCGTGAACCATCGAGTTTTTGAACGCAAGTTGCGCCCGAAGCCATTAGGCCGAGGGCACGTCTGCCTGGGCG
TCCCACAATGCGTCGCTCCCCCACTCACCCCAGGGTGTCGTGTGAGGTGCGGATATTGGCTCCCCGTGTGCTATCGTGCGC
GGTCAGCCTAAAAGTGAATCCCGACGATCGGTGTCGCGGCAAGTGGTGGTTTCCAAACCGTTGCATGTTGTCGTGCGCGCT
TCTATCGCCCTCGGTGACTCCTTGACCCTGAAGCTCCATTAGAATGGTGCCACGATCGAG

4. OBJETIVOS (ALCANCES):

Generales

- Diseñar y entrenar un Algoritmo Genético para realizar la alineación génica en forma eficiente.
- Presentar una aplicación de Algoritmos Genéticos.
- Crear una herramienta útil para el área de Sistemática Filogenética

Particulares

- Identificar el problema.
- Lograr un mecanismo eficiente para la alineación de secuencias génicas.
- Diseñar un Algoritmo Genético y programarlo.
- Implementar el criterio de eventos para seleccionar alineamientos útiles.
- Mejorar el tiempo de alineación con respecto a otros métodos o productos.
- Publicar un artículo con los resultados.

5. METODOLOGÍA UTILIZADA

- Revisar y analizar el estado del arte del problema.
- Identificación de puntos críticos en la alineación de secuencias.
- Diseño del Algoritmo Genético.
- Forma de representar las soluciones y el espacio de búsqueda.
- Forma de representar el mecanismo de evaluación.
- Forma de representar la función objetivo.

- Implementación del algoritmo genético.
- Entrenamiento del algoritmo.
- Fase de pruebas con un número especifico reducido de secuencias.
- Experimentación y Análisis.
- Presentar resultados.
- Conclusiones y conocimiento nuevo.
- Escritura de tesis.

DESGLOSE DE CAPITULOS.

Introducción

Capitulo I
- Introducción a la sistemática filogenética.
- Uso de secuencias de ADN en la sistemática filogenética.

Capitulo II
- Introducción a Algoritmos Genéticos.
- Historia
- Como funcionan los Algoritmos Genéticos

Capitulo III
- Métodos actuales de alineación.
- Como funcionan los métodos actuales.
- Problema actual de los métodos utilizados

Capitulo IV
- Diseño del algoritmo.
- Desarrollo
- Implementación
- Entrenamiento y pruebas
- Interfase
- Análisis de resultados contra otros métodos.

Conclusiones

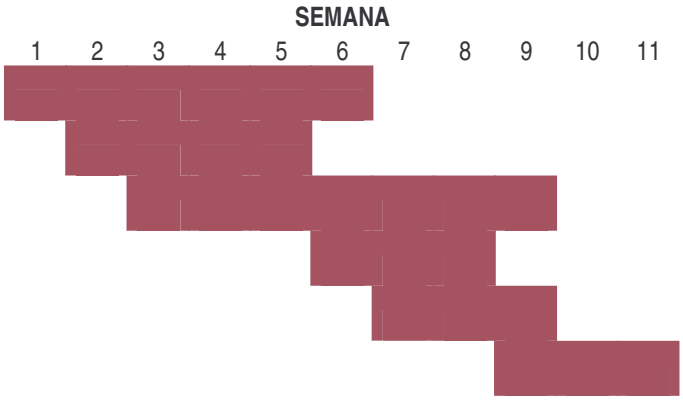
Referencias

6. CRONOGRAMA

TRIMESTRE I

Actividades

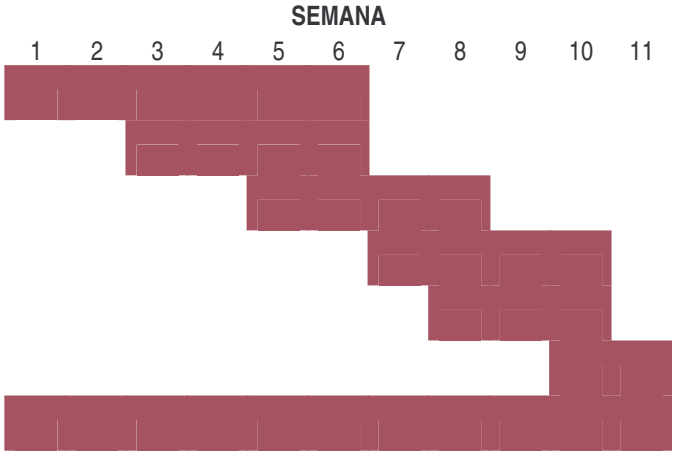
- Revisar y analizar el estado del arte del problema.
- Identificación de puntos críticos en la alineación de secuencias.
- Diseño del Algoritmo Genético.
-
- Forma de representar las soluciones y el espacio de búsqueda.
- Forma de representar el mecanismo de evaluación.
- Forma de representar la función objetivo.
-



TRIMESTRE II

Actividades

- Implementación del algoritmo genético.
-
- Entrenamiento del algoritmo.
-
- Fase de pruebas con un número especifico reducido de secuencias.
- Experimentación y Análisis.
-
- Presentar resultados.
-
- Conclusiones y conocimiento nuevo.
-
- Escritura de tesis.
-



7. RECURSOS.

REQUERIDOS

- Secuencias genéticas para realizar la alineación.
- Software. Programa para desarrollo y sistema operativo con licencia.
- Hardware. Equivalente a PIV, con periféricos normales.

DISPONIBLES

- Secuencias genéticas para realizar la alineación.
- Software. Programa para desarrollo y sistema operativo con licencia.
- Hardware. Equivalente a PIV, con periféricos normales.

8. BIBLIOGRAFÍA, REFERENCIAS y FUENTES.

Libro: Autor, título del libro, editorial, lugar de publicación, año.
Artículo: Autor, Título del trabajo, nombre de la revista Vol. xx No.xxx, fecha, pp.

[1] David E. Goldberg. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Publising Company. United Stated of America. 1989.
[2] Zbigniew Michalewicz. Genetic Algorithms + Data Structures = Evolution Programs
[4] H. Ochotorena. Apuntes del curso de Sistemática Filogenética. Posgrado en Ciencias Biológicas-UNAM. Mexico. Inédito.
[21] Javier Ramírez Rodríguez. Complejidad computacional de algoritmos. Ed. UAM-A. México. 1993.
[22] John H. Holland. Adaptation in Natural and Artificial Systems: An Introductory Analysis With Applications to Biology, Control, and Artificial Intelligence (Complex A)

[3] H. Ochoterena. Independence of alignment and phylogenetic reconstruction and their optimality criteria. Cladistics Vol.. 20, 2004, 91. (Resúmenes de la XXII reunión anual de la sociedad Willi Hennig, NY, EUA, 2003).
[5] Luke Sheneman and James A. Foster. Envolving Better Multiple Sequence Alignments. IBEST. 2004.
[6] A.E. Ruano, P.J. Fleming, C. Teixeira, K. Rodriguez Vazquez, C.M. Fonseca. Nonlinear identification of aircraft gas-turbine dynamics. Elsevier. 2003
[7] Cedric Notredame and Desmond G. Higgins. SAGA: sequence alignment bye genetic algorithm. Oxford University Press. 1996. 1515-1524.
[9] Matin Vingron y Michael S. Waterman. Sequence alignment and penalty choice. Academic Press Limited. 1994. 1-12
[10] Burkhard Morgenstern, Kornelie Frech, Andreas Dress, Thomas Werner. DIALING: Finding local similarities by multiple sequence alignment. Bioinformatics Vol. 14 No. 3. 1998. 290-294.

[8] Mark E. Siddall. Multiple Alignment. http://research.ammnh.org/users/siddall/methods/align.html
[11] http://es.wikipedia.org
[12] http://www.neurocomputing.org/index.html
[13] http://www.icpress.co.uk/
[14] http://www.iscb.org
[15] http://mitpress.mit.edu
[16] http://citeseer.ist.psu.edu/cs
[17] http://uxdea4.iimas.unam.mx/gcb/en/default.htm
[18] http://www.ieee.org
[19] http://www.elsevier.com/wps/find/homepage.cws_home
[20] http://research.amnh.org/users/siddall/methods/align.html