INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

Generación de Árboles Filogenéticos por medio de Algoritmos Genéticos de Función Objetivo Híbrido.

TESIS QUE PARA OPTAR EL GRADO DE
MAESTRO EN CIENCIAS COMPUTACIONALES
PRESENTA

Ryosuke Luis Alberto Watanabe Sakurazawa

Asesor: Dr. Edgar Vallejo Clemente

Jurado:	Dr. FERNANDO RAMOS,	Presidente
	Dr. EDGAR VALLEJO CLEMENTE,	Secretario
	Dr. VICTOR DE LA CUEVA,	Vocal

Atizapán de Zaragoza, Edo. Méx., Septiembre de 2002.

INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

Generación de Árboles Filogenéticos por medio de Algoritmos Genéticos de Función Objetivo Híbrido.

TESIS QUE PRESENTA

RYOSUKE LUIS ALBERTO WATANABE SAKURAZAWA

MAESTRÍA EN CIENCIAS COMPUTACIONALES
MCCI 00

AGRADECIMIENTOS:

A mi asesor DCC. Edgar Vallejo por creer en mí cuando yo dudaba y darme la oportunidad abriéndome las puertas para lograr mis metas. Por ser mi pilar de apoyo.

A mi Padre por exigirme ser la persona que soy.

A mi Madre por apoyarme sin cuestionar.

A mi novia Judith por todo su apoyo y comprensión que hizo que no claudicara. Albert Einstein tenía razón la energía más fuerte es la del amor.

Al DBM. Enrique Morett por darme la seguridad.

Al DCC. Roberto Gómez por su apoyo moral y dirigirme.

Al DCC. Fernando Ramos por dejarme pertenecer al grupo más importante para mí.

Al DCC Victor de la Cueva quién evaluó mi trabajo y me motivó.

A todo el equipo de WINTER que me apoyó para esta investigación.

A todos ellos les agradezco desde el fondo de mi alma porque esta meta no hubiera sido posible sin el apoyo de todos ellos. Les doy las gracias desde el fondo de mi ser.

ÍNDICE

RESUMEN:

Este documento propone un nuevo método para el estudio filogenético. Este método propuesto utiliza algoritmos genéticos con una fución objetivo híbrida para la evaluación y generación de árboles filogenéticos.

Para la generación de los árboles filogenéticos, el algoritmo genético, crea una población de árboles aleatoria cuyas hojas o nodos son diferentes. Cada elemento (árbol) se evalúa mediante una función objetivo híbrida, la cual combina dos criterios normalizados: máxima parsimonia y matríz de distancia (DMM) que darán su valor de aptitud a cada elemento, posteriormente se realizan las operaciones básicas del algoritmo genético como son la selección, cruza y mutación.

La hipótesis en la que se fundamenta este trabajo de investigación es que las formas de evaluación existentes para generar árboles filogenéticos no siempre son el mejor modelo para generar el árbol. y que la evaluación híbrida es una forma de aprovechar las fortalezas y las minimizar debilidades de ambos métodos.

Para validar su desempeño, se realizaron pruebas con cadenas ADN de HIV (virus del SIDA) y con cadenas de proteínas de E-Coli (Escherichia Coli). Estos árboles se comprobarón con los obtenidos con los métodos de Matriz de Distancia y Máxima Parsimonia. También los resultados se interpretarón y evaluarón por expertos de (UNAM) IBT (Instituto de Biotecnología de la Universidad Autónoma Nacional de México).

Los resultados experimentales obtenidos indican que el método propuesto es capaz de inducir relaciones filogenéticas significativas dentro de una misma especie.

Lista de Tablas

Lista de Figuras

1. INTRODUCCIÓN:

1.1.- La Bioinformática:

Los recientes avances en los métodos experimentales de la biología molecular y las nuevas tecnologías de la información condicionaron el surgimiento de una disciplina que generó vínculos indisolubles entre la Informática y las Ciencias Biológicas: la Bioinformática.

Ella se encuentra en la intersección de las ciencias de la vida con las ciencias de la información. Es un campo científico interdisciplinario que se propone la investigación y el desarrollo de sistemas computacionales que faciliten la comprensión del flujo de información desde los genes a las estructuras moleculares, su función bioquímica, su conducta fisiológica y finalmente, su influencia en las enfermedades y la salud.

Entre los principales factores que han favorecido el desarrollo de esta disciplina, se encuentra el impresionante volumen de datos sobre secuencias generadas por los distintos proyectos genoma (tanto el humano como el de otros organismos); los nuevos enfoques experimentales basados en biochips que permiten obtener datos genéticos a gran velocidad, bien de genomas individuales (mutaciones, polimorfismos), o de enfoques celulares (expresión génica); así como el desarrollo de Internet y la World Wide Web, que permite el acceso mundial a las bases de datos de información biológica.

El término Bioinformática es relativamente reciente, y apareció en la literatura a principios de 1990, cuando comenzaba a estructurarse el llamado "Proyecto Genoma Humano" y el “National Center for Biotechnology Information” de los Estados Unidos, daba sus primeros pasos. En los primerras etapas, su vinculación con la Informática Médica (IM) se debió solamente a la similitud sintáctica, así como al indispensable uso de las computadoras por parte de ambas disciplinas.

La Bioinformática, por su parte, tiene como reto principal ofrecer una respuesta a la avalancha de datos procedentes de la Genómica. Mientras, que hace unos años, los resultados de

los experimentos podían interpretarse sobre el cuaderno de laboratorio, hoy se necesitan bases de datos y técnicas de visualización sólo para almacenarlos y comenzar a estudiarlos. Ella evolucionó desde un conjunto de técnicas hacia una verdadera ciencia, al aportar el componente de análisis para entender la genómica e integrar sus datos que permitieran crear modelos predictivos para los sistemas biológicos.

No obstante, con el advenimiento del nuevo milenio, el procesamiento de la información genética (bioinformática) y de la información clínica (informática médica) amenaza con fundir ambas disciplinas en una sola, que algunos autores han definido como Informática Biomédica. En la medida que se genera información sobre el genoma humano y ésta se vincula con el conocimiento médico de las enfermedades, esta definición ha comenzado a hacerse realidad. Los datos que maneja la bioinformática tienen cada vez más presencia en la práctica médica; por tanto, conseguir la unificación de la información clínica con la información molecular representa el desafío más importante de esta disciplina durante el presente siglo.

Sus principales aplicaciones, según los resultados obtenidos por estadísticas de la Medline, fueron la gestión de datos en los laboratorios, la automatización de experimentos, el ensamblaje de secuencias contiguas, la predicción de dominios funcionales en secuencias génicas, la alineación de secuencias, las búsquedas en bases de datos de estructuras, la determinación y predicción de la estructura de las macromoléculas, la evolución molecular y los árboles filogenéticos. Las especialidades médicas que recibieron una mayor influencia de la Bioinformática fueron la Genética Médica, la Bioquímica Clínica, la Farmacología, las Neurociencias, la Estadística Médica, la Inmunología, la Fisiología , la Oncología, Epidemiología, Filogenética y Secuenciación. [Medline, 2004]

1.2.- Planteamiento del problema:

La filogenética molecular estudia la relación entre los seres vivos por medio de árboles genealógicos entre las especies del reino natural, y utiliza mecanismos a nivel molecular, tales como el ADN o proteínas.

Los árboles filogenéticos son diagramas de relación que muestran la proximidad entre las especies en cuanto a la cercanía de sus genes. La construcción de estos modelos se basa en diferentes métodos de evaluación de las posibles soluciones. Por ejemplo, el método de Matríz de Distancia (DMM) toma la diferencia entre las secuencias comparadas, mientras que Máxima Parsimonia busca puntos de referencia donde se encuentre una evolución homogénea.

Cada método tiene sus particularidades donde explota una característica para evaluar la relación entre las especies, es decir, la relación a nivel molecular (ADN,ARN o Proteínas). Sin embargo, ninguno de los métodos es aceptado en su totalidad debido a que una fortaleza atrae una debilidad, específicamente en el caso de los métodos de DMM y Máxima Parsimonia. [Gaur et al, 2000]

Todos los métodos de evaluación se reducen a un simple término: "Cambio evolutivo". La mutación es la clave de la evolución, debido a que por medio de prueba y error, el más apto sobrevive. Reflejando esto en la relación entre especies (evolución), es la diferencia entre los organismos, y reflejado en un nivel de filogenética moleculár (árboles filogenéticos); es la relación entre ADN o Proteínas, en cuanto a la similitud entre varias secuencias que se comparán. Es decir, comparación entre la diferencia o similitud entre las secuencias, en otras palabras, mutación.

Sin embargo, ningún método es completo ya que la mutación es considerada en su totalidad o ignorada casi por completo al momento de comparar las secuencias. En términos prácticos, esto implica que cuando se compara la totalidad de las secuencias se está contando toda la mutación o errores y por el otro lado, cuando se comparan solo los puntos significativos de las secuencias, se está ignorando mucha información que posiblemente contenía datos claves. Esto produce árboles

que no siempre son los mejores, ya que dependen de la longitud, la similitud y los patrones de agrupamiento o zonas donde se pueda localizar un patrón.

La sistemática cladística (cladismo o sistemática filogenética) tuvo su origen en el libro publicado en 1950 por Willi Hennig (traducción castellana de 1968). Desde entonces su popularidad ha ido en aumento y con los años el cladismo ha ido mucho más allá del contexto planteado por Hennig originalmente. La meta del cladismo es producir hipótesis "comprobables" de las relaciones genealógicas entre grupos monofiléticos de organismos. Como metodología, está basado completamente sobre la "descendencia común", o sea, la genealogía estricta. El dendrograma usado por los cladistas es denominado cladograma o árbol y está construido, únicamente, para mostrar la genealogía, es decir, las relaciones ancestral-descendiente.

La Fenética, denominada a veces también como Taxonomía numérica o Taximetría, se desarrolló como consecuencia de la disponibilidad creciente de computadores al final de los años 50. La base filosófica para la "fenética numérica" es el argumento de que como nunca es posible conocer con certeza cuál de las diferentes filogenias en competencia es la correcta, la descripción real de los individuos de un grupo ayudará a conocer la evolución de ese grupo, pero nada más. Por lo anterior, los organismos deberían ser clasificados estrictamente en función de la conveniencia, en lugar de establecer clasificaciones basadas sobre reconstrucciones hipotéticas de la historia filogenética de un grupo de animales o plantas, o sea, como los libros en una biblioteca.

1.3.- Hipótesis:

La hipótesis de esta investigación establece que la aplicación de los algoritmos genéticos, basados en más de un criterio de evaluación para la construcción de filogenias produce resultados que son un punto de equilibrio entre los métodos cladistas y fenetistas, es decir, un árbol filogenético balanceado en cuanto a la estructura y la similitud entre los elementos comparados logrando una clasificación más eficiente de la relación entre las especies, además de mejorar la velocidad de procesamiento con respecto a los métodos tradicionales.

Para poder definir la relación que existe entre las especies o cadenas a comprar, la mutación juega un papel importante. Es sabido que alrededor del 2.4% de la información genética tiene una mutación a nivel genético, pero esta mutación se puede originar no sólo cuando se genera un nuevo organismo, sino también cuando el ADN se transcribe o el ARN se traduce. Asimismo, el ADN sufre trastornos a lo largo de la vida de un organismo, siendo este también una mutación. [Kimura,1983]

Teniendo en cuenta todos estos aspectos de la mutación, es imposible ignorar completamente la mutación en todos los aminoácidos o proteínas, pero también no es certero tomar en cuenta la mutación en su totalidad, debido a que puede ser sólo un cambio local que no tiene mucho significado y que no permitirá fijar una relación entre las especies. Es decir, una mutación neutral [Kimura, 1983]]

Para poder equilibrar esta evaluación y poder tomar en cuenta la mutación y también tomar en cuenta el patrón de similitud y minimizar el efecto negativo de la mutación se propone hibridizar la función objetivo para la generación de árboles filogenéticos.

El valor de normalización para decidir cuánto tomar en cuenta el valor de la mutación, se utiliza la teoría neutralista de la evolución molecular. Fijando así al valor de normalización para asignarle a DMM un valor de 2.4 % que es el que se establece como valor promedio de evolución molecular para los ácidos nucleicos y 0.8 % para las proteínas. [Kimura,1983]

Estos valores se basan en estudios realizados por Motoo Kimura, y en su opinión, la mayoría de los genes mutantes son selectivamente neutros, es decir, no tienen selectivamente ni más ni menos ventaja que los genes a los que sustituyen; en el nivel molecular, la mayoría de los cambios evolutivos se deben a la deriva genética de genes mutantes selectivamente equivalentes.

Existe también otro problema al momento de generar los árboles filogenéticos, el cual se refiere al gran número de combinaciones que se requiere para poder encontrar el mejor. Lo cual hace de este un problema con espacio de búsqueda muy grande y de ahí que sea necesaria la utilización de heurísticas que permitan llegar a una solución aceptable sin tener que realizar cálculos exhaustivos.

El algoritmo genético es un procedimiento de búsqueda utilizado para poder obtener resultados de manera rápida y efectiva, ya que se sabe que el algoritmo genético es una heurística que produce buenos resultados en una amplia variedad de aplicaciones. El algoritmo genético opera simulando el modelo evolutivo Darwiniano de la supervivencia del más apto utilizando la mutación, reproducción y selección por médio de una función de aptitud o función objetivo. La función objetivo de los algoritmos genéticos se expresa generalmente en forma de ecuaciones que pueden ser simples o compuestos, es decir, de un solo objetivo o multiobjetivo, así mismo se pude tener una ecuación formada a partir de dos o más ecuaciones a lo cual llamamos función objetivo híbrida. [Martí,2000]

El uso del algoritmo genético para la generación de árboles filogenéticos ha sido previamente demostrado por Bates (2000) , donde utiliza los algoritmos genéticos con una función objetivo basada en parsimonia para generar los árboles filogenéticos. De esta manera, comprueba la efectividad de los algoritmos genéticos en cuanto a la velocidad y funcionalidad para obtener árboles válidos en un tiempo razonable. Este trabajo extiende la utilización del algoritmo genético en la generación de árboles filogenéticos pero con una función objetivo híbrida debido a que creemos que la unión de dos métodos proporciona un enfoque más completo.

Una vez realizada esta hibridación, se verifica el funcionamiento del sistema, tomando en cuenta dos tipos de información. La primera de secuencias ADN de HIV tipo 1 y 2 (americano,africano) y la segunda con secuencias de Proteínas de E-Coli.

1.4.- Objetivos:

El objetivo de esta investigación es crear una metodología híbrida para generar árboles filogenéticos utilizando algoritmos genéticos. Adicionalmente, se pretende realizar una validación para verificar el correcto funcionamiento de dicho método.

Específicamente, deseamos lograr combinar varios métodos de generación de árboles filogenéticos para suplir la deficiencia de un método con la fortaleza de otro, en este caso particular con los métodos de matriz de distancia y con máxima parsimonia, aunque esto será la base para poder hibridizar varios métodos que se complementarán y formarán un nueva forma de generación de árboles filogenéticos.

1.5.- Contribuciones esperadas:

Esperamos que utilizando la implementación de algoritmos genéticos con función objetivo híbrida para la generación de árboles filogenéticos, el tiempo de respuesta para generar árboles filogenéticos será más rápido que los métodos convencionales. Asimismo, por medio de la utilización de una función objetivo híbrida esperamos observar resultado novedosos, generando un árbol que sea una combinación de los dos métodos y que tenga una estructura similar a los métodos utilizados, así como la configuración del árbol deberá cumplir con la lógica de las especies comparadas.

Pretendemos que la forma de clasificación y el árbol generado sea más cercana a la relación entre la especie, debido a que se considerará no solo la similitud sino que también se tomará en cuenta los puntos de evolución, lo cual permitirá generar un árbol más preciso por la evaluación de máxima parsimonia que genera grupos, lo que permite observar puntos de cambio o sitios de cambio evolutivo y con DMM se verá la relación de mutación que con la unión de estos dos métodos se toma en cuenta en su totalidad los aspectos de la evolución y mutación.

Al considerar más de un criterio de evaluación tomada en un mismo cálculo, se espera obtener resultados que sean mejores que solo utilizando un criterio de evaluación, esto se fundamenta en la complementación de métodos a utilizar.

Este documento de tesis se encuentra organizado en la forma siguiente. En el capitulo dos se explica brevemente el marco teórico a cerca de la evolución molecular, la biología molecular y los métodos de construcción de filogenias. Posteriormente, se explican los algoritmos evolutivos en especial, los algoritmos genéticos, la programación genética y la construcción de filogenias utilizando algoritmos genéticos.

En el capitulo 3 se explica el modelo propuesto para la generación de árboles filogenéticos utilizando algoritmos genéticos con función objetivo híbrida, los criterios de evaluación y las métricas utilizadas.

En el capitulo 4 se explican los experimentos y resultados realizados. Para terminar con las conclusiones en el capitulo 6 que incluye trabajos futuros.

Por último tenemos un glosario de términos el cual contiene definiciones de términos biológicos y computacionales para facilitar el entendimiento al lector.

2 Marco Teórico

2.1 La Evolución Molecular.

La evolución molecular, definida por algunas personas como el reloj de la vida, estudia cómo las moléculas cambian a través del tiempo evolutivo. Esta evolución se puede observar como cambios de nucleótidos en el ADN y en aminoácidos de las proteínas codificadas. Los biólogos evolucionistas moleculares, buscan encontrar el mecanismo involucrado en la evolución de las moléculas y dónde se observan los cambios modelados por la selección natural o la deriva genética.

2.1.1 La Biología Molecular

Actualmente, el término biología molecular es tratado frecuentemente en la rama científica médica, biológica y genética. Este término se refiere a la biología de moléculas relacionadas a los genes y a sus productos, así como a la herencia. Esta área de estudio también se denomina genética molecular, que estudia la evolución molecular.

Esta ciencia tuvo su origen en los comienzos de 1800, cuando algunos curiosos empezaron a preguntarse sobre los patrones de la herencia, tales como el color del cabello o de los ojos. Pero no fue hasta la Segunda Guerra Mundial cuando se empezó a relacionar estos fenómenos con procesos fundamentados en la herencia molecular [Clark & Russell, 2000].

De esta manera, la genética molecular empieza sus estudios de la genética por parte de la biología molecular.

2.1.2 La Evolución Molecular

El nacimiento de la genética moderna se debe al descubrimiento de Gregor Mendel, cuyo logro fue tomar a cada característica hereditaria como una unidad y no todas en conjunto. (Mendel descubrió las bases de las leyes de la genética realizando cruzas entre las plantas). En nuestros días, cada característica examinada por Mendel es determinada como un gen singular.

Los genes son unidades genéticas de información, cada uno provee de instrucciones para alguna característica del organismo. Cada gen puede existir en formas alternativas, por ejemplo, flores blancas o rojas, a estos genes que especifican los cambios se les denomina alelos. Nótese que diferentes alelos del mismo gen están relacionados, pero una variación menor a nivel molecular produce diferentes expresiones en el ser viviente tales como el color.

El gen determina cuándo la flor será roja o blanca, esto debido a las reacciones biosintéticas ocasionadas por proteínas especiales conocidas como enzimas. Cada enzima es capaz de generar una reacción química particular. Como se puede observar en la figura

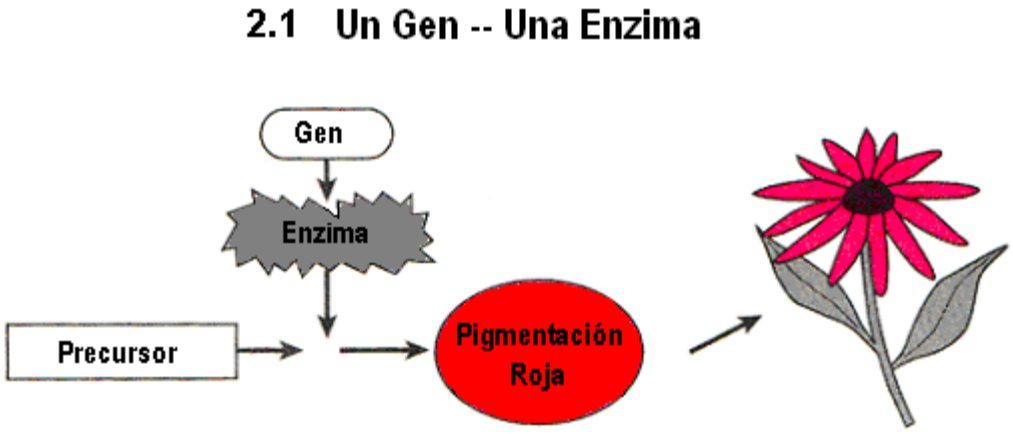


Figura 2.1 Un Gen que reacciona con una enzima y con la unión de un precursor quien pone la pauta de la información genética, reacciona para producir una pigmentación roja en la flor.

Las relaciones químicas están relacionadas en cada gen que se basa en la descripción genética que es llamada genotipo y la descripción visible en el organismo que es llamada fenotipo.

El segundo descubrimiento importante fue el ADN (ácido disoxiribonucleico), el cual es el constitutivo de la información genética. Cada parte del ADN se encuentra organizado en cromosomas que son grandes moléculas de ADN conformadas por ácidos nucléicos de secuencias de bases. Existen cuatro tipos de bases, la Adenina denominada por la letra "A", la Guanina denominada por "G", Citosina denominada por "C" y la Timina denominada por la letra "T". [Clark & Rusell, 2000]

El genoma consiste de una colección de cromosomas, los cuales describen la estructura y características de un ser viviente.

Dependiendo del tipo del organismo, existen diferencias en las cadenas de ADN. Estas puede ser haploides, los cuales sólo poseen una copia de la información genética o diploides si poseen dos copias; la mayoría de los organismos pluricelulares o eucariotes poseen cromosomas diploides.

En sentido amplio, se entiende por evolución, los cambios que se dan en las características de los organismos en periodos largos de tiempo. Esta definición incluye la dimensión temporal y la noción de cambio. Este cambio tiene que afectar la información genética, porque son únicamente las características con base genética las que se transmiten de una generación a la siguiente. Cuando en lugar de hablar de evolución en general se habla de evolución molecular, se hace referencia a la evolución de las moléculas, y más concretamente, a la evolución de los ácidos nucleicos (que constituyen el material hereditario, y por tanto los genes) y de las proteínas (que son el producto primario de la expresión de estos genes).

Comparando la secuencia de aminoácidos de una determinada proteína en especies diferentes, o entre individuos, se detectan alteraciones en los aminoácidos como se puede observar en la figura 2.2:

Figura 2.2 Comparación de Secuencias

Nombre	Sequencia														
Alpha	A	A	C	G	U	G	G	C	C	A	-	A	A	U	
Beta	-	-	G	-	-	-	-	-	-	-	-	-	-	C	
Gamma	-	-	C	-	U	U	-	C	-	U	-	-	C	A	
Delta	G	G	U	A	-	U	U	-	G	G	-	C	C	-	

Comparación de secuencias, se alinean entre ellas y se buscan sitios iguales.

Lo mismo sucede si comparan fragmentos concretos de ADN. La constatación de que a nivel molecular hay diferencias interespecíficas (divergencia) e intraespecifica (polimorfismo) hace que nos preguntemos por el motivo de los cambios observados y que intentemos distingur qué factores evolutivos son responsables. Entre estos factores se encuentran la selección natural, que da lugar a las adaptaciones, y la deriva genética, que provoca cambios aleatorios no adaptables.

Las diferencias que se detectan al comparar las secuencias de un gen determinado entre dos o más especies, o bien entre individuos de una misma especie, constituyen una fracción pequeña de todos los cambios (mutaciones), que es el destino evolutivo de las mutaciones.

La impresión que muchos tenemos a cerca de los mutantes es lo que la ciencia-ficción nos ha mostrado: seres dotados de fuerzas sobrenaturales o monstruos horribles, sin embargo eso no es la realidad, la realidad es que: "Todos somos mutantes" (Clark & Russell,2000:153).

La naturaleza no es perfecta, sino que muchos procesos de la biología molecular están sujetos a "errores". La equivocación en el material genético de una célula, es conocido como mutación.

A nivel molecular, la mutación es la alteración en las moléculas de ADN. Por esta razón, la mutación puede pasar de células de los padres a sus descendientes. A estos les llamamos defectos heredados.

Pero entonces, ¿por qué no estamos muertos o deformados? La respuesta es que existen diferentes tipos de mutaciones y que la mayoría solo tiene un efecto menor, de hecho, muchos no parecen generar defectos. Relativamente son pocas las mutaciones que causan un cambio tal que logre llamar la atención. Por otra parte, la mayoría de los organismos pluricelulares eucariotes poseen dos copias de cada gen debido a que nuestro ADN es diploide. Esto significa que si una copia sufre una mutación, existe una copia de "respaldo" que puede ser expresada.

2.1.2.1 Alteración del ADN por mutación.

Tomando en cuenta que la mutación es un cambio en la secuencia base del ADN, existen muchas posibilidades de cambios que pueden suceder. Los tipos de cambio son [Gaur,2000]:

a) Sustitución: Cuando un ácido nucleico es cambiado por otro diferente.
b) Inserción: Cuando uno o varios ácidos nucleicos es introducción en algún lugar de la secuencia del ADN.
c) Borrado: Cuando se elimina uno o varios ácidos nucleicos dentro del ADN.
d) Inversión: Cuando se cambia el orden de los ácidos nucleicos de un segmento dentro del ADN.

Estos errores pueden ocurrir en dos partes. En la primera, si la molécula de ADN se replica y transmite el error a la segunda generación. La segunda, si se realiza la transcripción para copiar al ARN. Esto hace que el error pase a la cadena de ARN y lo altera.

2.1.2.2 Tipos de mutación según sus efectos.

Existen diferentes tipos de mutación según el efecto que ocasionan, unos pueden ser silenciosos, que es la mutación cuyo efecto no altera el funcionamiento correcto de la célula; esto puede ocurrir debido a dos consecuencias. La primera es que la combinación de amino ácidos pueden producir proteínas similares, por lo cual si la combinación es alterada pero produce la misma proteína, el efecto es nulo. La otra causa es debido a que en el ADN existen partes que codifican a la proteína llamados exones y partes que no codifican a la proteína llamadas intrones, si la mutación se presenta en el intrón del ADN, entonces el efecto también es nulo.

El otro tipo de mutación es el que ocasiona errores en el amino ácido, éste sucede cuando una secuencia del ADN resulta en remplazar un amino ácido por otro para codificar la proteína. Es entonces que se altera el funcionamiento de la célula ocasionando mutaciones visibles o con efecto dañino.

El tercer tipo de mutación es otra mutación dañina, la cual hace que se generen codones de paro (codón de paro es el un conjunto de tres aminoácidos que establecen donde termina una secuencia determinada) que son las regiones del ADN donde se puede adherir un mARN (ARN Mensajero) para segmentar y sólo producir la proteína específica, esto significa que al ocasionarse una mutación en una zona, ésta cambia por UAA, UAG o UGA que son cadenas de paro para la codificación de una proteína, ocasionando con esto que no se genere el resto de la proteína que se debería de generar.

Un aspecto importante que hay que tomar en cuenta es que no toda mutación define a una nueva especie, esto se debe a los tipos de mutación que se mencionarón arriba. Esto se descubrió en 1980, cuando dos genes de la hemoglobina fueron secuenciados. Aunque ambos codificaban el mismo producto, sus secuencias de nucleótidos diferían en el 0.8% si sólo se tenían en cuenta las sustituciones de un aminoácido por otro, y en un 2.4% si se incluían en la comparación los aminoácidos presentes en un gen y ausentes en otro. Otros genes secuenciados posteriormente en otros organismos llevaban a la misma conclusión: en la secuencia de ADN, los organismos quizá sean heterocigotos para todos sus loci. [KIMURA,1983]

La gran variación revelada por estos estudios constituye uno de los fundamentos de la teoría neutralista, otro de los desafíos a la teoría sintética de la teoría propuesta por Darwin.

Su principal expositor es Motoo Kimura, y en su opinión, la mayoría de los genes mutantes son selectivamente neutros, es decir, no tienen selectivamente ni más ni menos ventaja que los genes a los que sustituyen; en el nivel molecular, la mayoría de los cambios evolutivos se deben a la deriva genética de genes mutantes selectivamente equivalentes. (La deriva genética consiste en el cambio puramente aleatorio de las frecuencias génicas, debido a que cualquier población consta de un número finito de individuos. La razón es la misma por la que es posible que salga cara más de 50 veces cuando lanzamos una moneda al aire cien veces). En otras palabras, esto significa que un gen mutado tiene la misma probabilidad de transmitirse a la siguiente generación que el gen no mutado.

De esta manera se llega a la conclusión de que solo un 2.4 % aproximadamente de las mutaciones, sirven como una fijación el cual sea un elemento de evolución. [KIMURA,1983]

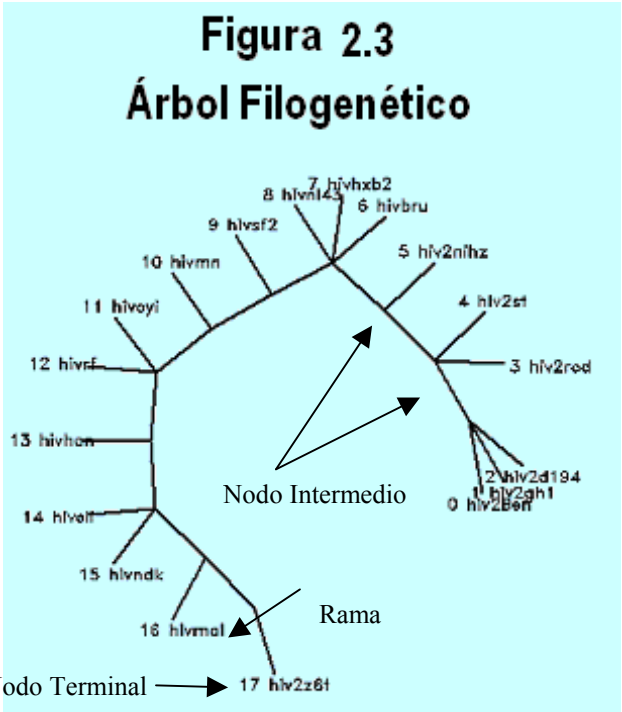
2.1.3 Métodos de construcción de filogenias

Los árboles filogenéticos son diagramas de relación que muestran la proximidad entre las especies en relación a la cercanía de sus genes por medio de diferentes métodos de evaluación de diferentes soluciones posibles.

En cada árbol la proximidad entre nodos la señalan las ramas. Cada rama está unida por medio de otras ramas y en su punto final tiene un nodo terminal que contendrá a una especie y la distancia que exista entre ellas. Las ramas también están unidas a nodos intermedios que son la relación en donde se separan las especies.

Los árboles pueden ser enraizados y desenraizados, donde los enraizados son los que tienen una raíz, es decir, se sabe que especie o elemento es el antecesor y cuales son los predecesores en el tiempo. Por otra parte, en los árboles desenraizados, sólo se sabe su relación pero no en cuanto a la evolución en el tiempo. (ver fig. 2.3)

Figura 2.3
Árbol Filogenético



La construcción de árboles filogenéticos puede realizarse a través de diferentes métodos, los cuales poseen tanto sus fortalezas como sus debilidades. El utilizar una combinación de diferentes métodos se propone en este trabajo como un medio de explotar las ventajas de un método para suplir las debilidades de otro método.

Los métodos más utilizados en la práctica son: DMM (Distance Matrix Méthod) y Máxima Parsimonia, los cuales se explicarán posteriormente. [Graur y Li, 2000]

Para todos los métodos, es necesario realizar un conjunto de combinaciones muy numerosas para poder lograr comparar diferentes árboles alternativos y muchas veces es necesario generar la totalidad de combinaciones de árboles para evaluarlos. Este proceso constituye un cálculo muy complejo para la mayoría de los equipos de cómputo disponibles en la actualidad. Específicamente, el orden de crecimiento está dado por:

- Número de árboles bifurcados desenraizados :

$$NN = \frac{(2n-5)!}{[2n-3(n-3)!]}$$

Donde n es el número de nodos y NN es el número de combinaciones de árboles desenraizados.

- Número de árboles bifurcados enraizados :

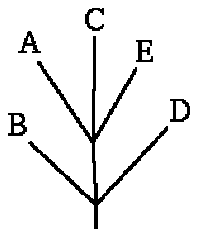
$$NE = \frac{(2n-3)!}{[2n-2(n-2)!]}$$

Donde n es el número de nodos y NE es el número de combinaciones de árboles enraizados.

Por ejemplo, un árbol con 10 nodos posee 86600951 árboles desenraizados y 4410806400 enraizados.

2.1.3.1 Formato Newick

El estándar Newick para representar árboles en un formato adecuado para computadoras, fue creado en 1857 por el famoso matemático inglés Arthur Cayley. Si tenemos el árbol enraizado:



Entonces el árbol es representado por la secuencia imprimible de caracteres:
(B,(A,C,E),D);

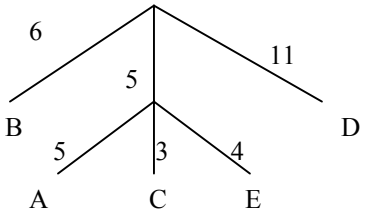
El árbol termina con punto y coma. El nodo inferior en este árbol es un nodo interior. Los nodos interiores son representados por pares de paréntesis. Entre ellos existe una representación de nodos que son descendientes inmediatos que son *B*, que es otro nodo interior y *D*. El otro nodo interior es representado por un par de paréntesis, las representaciones de sus descendientes inmediatos, *A*, *C* y *E*. En nuestro ejemplo, son terminales, pero en general, podrían ser nodos internos los cuales contendrían más paréntesis recursivamente.

En general:
Escribir *(*);*
Donde * es el par (nodo: tamaño); o *(* , *)*

Los nodos terminales son representados por sus nombres. Un nombre puede ser cualquier cadena de caracteres excepto espacios, punto y coma o paréntesis. En el caso de querer incluir espacios, se debe de utilizar el carácter "_". Cualquier árbol vacío como *(,(,,),);* es permitido. Los árboles pueden ser multiramificados en cualquier nivel.

Las ramas pueden ser incorporadas en un árbol escribiendo un número real, con o sin números decimales después de un nodo seguido de dos puntos. Este valor representará la longitud de la rama que se encuentre inmediatamente arriba del nodo. Y se puede representar utilizando el ejemplo anterior de la manera:

(B:6.0,(A:5.0,C:3.0,E:4.0):5.0,D:11.0);



Un árbol comienza en la primera línea del archivo en el que se almacene y puede continuar en las líneas subsecuentes. Se recomienda pasar al siguiente renglón cuando, después de escribir la coma que separa a dos ramas. Los espacios pueden ser agregados en cualquier punto excepto en medio del nombre de una especie o en la longitud de una rama.

El ejemplo anterior es una descripción de un subgrupo del estándar Newick. Los nodos interiores, por su parte, podrían contener nombres como se ejemplifica a continuación para su mejor entendimiento:

((mapache:19.19959,oso:6.80041):0.84600,((leon_marino:11.99700,
foca:12.00300):7.52973,((mono:100.85930,gato:47.14069):20.59201,
comadreja:18.87953):2.09460):3.87382,perro:25.46154);
(Bovino:0.69395,(Gibbon:0.36079,(Orangután:0.33636,(Gorila:0.17147,(Chimpance:0.19268,
Humano:0.11927):0.08386):0.06124):0.15057):0.54939,Ratón:1.21460);
(Bovine:0.69395,(Hylobates:0.36079,(Pongo:0.33636,(G._Gorilla:0.17147,
(P._paniscus:0.19268,H._sapiens:0.11927):0.08386):0.06124):0.15057):0.54939, Rodent:1.21460);
A;
((A,B),(C,D));
(Alpha,Beta,Gamma,Delta,,Epsilon,,,);

El estándar Newick no hace una representación única de un árbol, esto se debe a dos razones. Primero, el orden de derecha a izquierda de los descendientes de un nodo afecta la representación, aunque biológicamente sea irrelevante. Por lo cual *(A,(B,C),D);* sería igual a (A,(C,B),D);

Adicionalmente, el estándar representa un árbol enraizado. Para muchos propósitos biológicos, no es posible inferir la posición de la raíz. En este caso sería bueno poder representar árboles sin raíz. Y en este caso *(B,(A,D),C);* sería el mismo árbol sin raíz que *(A,(B,C),D); y ((A,D),(C,B));*

A pesar de estas limitaciones, la facilidad de representación y lectura-escritura del formato en programas de computadora lo ha transofmrado en un formato estándar *de facto*.

El estándar Newick fue adoptado en junio 26 de 1986 por un comité informal durante la junta en Durham, New Hampshire, de la Sociedad para el Estudio de la Evolución[4]. No existe aún una publicación formal del estándar Newick. Aunque Gary Olsen ha producido una descripción formal del estándar. [Archie, et al., 2000]

2.1.3.2 Métodos de Generación de Árboles Filogenéticos.

La inferencia de una filogenia es un proceso de estimación, en el cual el "mejor estimado" de una historia evolutiva está hecha con base en información incompleta. En el contexto de la filogenética molecular, usualmente no contamos con la información sobre el pasado; solamente tenemos información sobre las secuencias contemporáneas derivadas de organismos contemporáneos. Debido a que muchos árboles filogenéticos diferentes pueden ser producidos de cualquier grupo de OTU (Operational Taxonomical Unit), que es el valor que se utiliza para realizar la comparación entre especies, es necesario especificar criterios para seleccionar uno o algunos árboles que representen nuestra mejor estimación de una verdadera historia evolutiva. Muchos métodos de inferencia filogenética buscan cumplir esta meta al establecer un criterio de comparación entre filogenias alternativas y decidir cual árbol es mejor.

Entonces, una reconstrucción filogenética consta de dos pasos: El primero de la definición de criterio de optimalidad o función objetivo, el valor que es asignado al árbol y que es utilizado para comparar un árbol con otros. El segundo es el diseño de un algoritmo específico que calcule el valor de la función objetivo para identificar el árbol que tenga el mejor valor establecido para el criterio de mérito establecido.

4 La razón del nombre es que en la segunda y en la última sesión del comité se llevó a cabo en el restaurante Newick's en Dover y todos los integrantes disfrutaron de la langosta.

A través de los años, se han propuesto numerosos métodos de construcción de árboles. Para mayor detalle se les recomienda consultar Sneath and Sokal (1973), Nei(1987) y Felsenstein (1982,1988).

Una controversia que ha existido durante un largo tiempo es la disputa entre cladística y fenetística. La cladística se puede definir como el estudio del camino de la evolución. En otras palabras, los cladístas se interesan en preguntas como: ¿cuántas ramas existen entre un grupo de organismos?, ¿cuál rama conecta a cuál otra rama? y ¿cuál es el orden de la ramificación? El árbol que expresa a esta relación ancestro – descendiente es llamado cladograma. En otras palabras, un cladograma se refiere a la topología de un árbol filogenético enraizado.

Por el otro lado, fenetística es el estudio de la relación entre grupos de organismos en base al grado de similitud entre ellos ya sea molecular, fenotípica o anatómica. El árbol que expresa la relación fenetistica es llamado fenograma. Mientras un fenograma puede ser tomado como un indicador de relaciones cladísticas, los árboles no son necesariamente idénticos entre sí. Solo existe un caso en el cual ambos árboles serían idénticos, que es cuando existe una relación lineal en el tiempo de evolución y en el grado de la divergencia genética. Esto significa que la evolución esté siempre dada en un tiempo fijo.

Cada uno de las dos divergentes de la filogenética cuenta con su método dominante, para la cladística, el método utilizado principalmente es el de máxima parsimonia, mientras que para la fenetística, el método es el UPGMA (Unweighted Pair-Group Method with Arithmetic Means). [PEVZNER,2000]

2.1.3.3 Métodos de Matriz de Distancia.

En el método de matriz de distancia que es una forma cladista, la distancia evolutiva (usualmente es utilizando el número de sustituciones de nucleótidos o reemplazo de amino ácidos entre dos unidades taxonómicas) son calculadas para cada par de taxas, que es un grupo que contiene las características similares, y se construye el árbol filogenético utilizando un algoritmo basado en alguna relación funcional entre el valor de las distancias.

Existen varios métodos de construcción de árboles filogenéticos dentro de la forma Matriz de Distancias tales como UPMGA, *Transformed distance method, Sattath and Tversky's neighbors-relation method*, etc., sin embargo, solo trataremos UPMGA debido a que es el método más utilizados en la práctica.

2.1.3.4 Unweighted pair-group method with arithmetic means (UPMGA).

Este es el método más simple para reconstruir árboles filogenéticos que fue originalmente desarrollado para construir fenogramas taxonómicos, es decir, árboles que reflejan las similitudes fenotípicas entre OTUs (Socal y Michener, 1958). Sin embargo, puede ser utilizado también para construir árboles filogenéticos si el grado de evolución es próximo a ser constante entre las diferentes líneas de tiempo (periodo en el que evolucionan), de tal manera que existe una relación lineal aproximada entre las distancias de evolución y el tiempo de divergencia (Nei 1975). Para tal relación, deben de utilizarse métricas de distancia lineal, tales como el número de sustituciones de nucleótidos.

UPGMA emplea esencialmente un algoritmo de agrupación (clustering) secuencial, en el cual la relación entre la topología es identificada en relación a la similitud decreciente, y el árbol es generado de manera escalonada. En otras palabras, primero identificamos los dos OTU más parecidos entre sí y utilizar a estos como un solo OTU. Este nuevo OTU es llamado OTU compuesto. Para el nuevo grupo de OTUs, se realiza el cómputo de las nueva matriz de distancia y se identifica el siguiente par con mayor similitud. Este proceso se repite hasta que sólo quedan dos OTU.

Para ilustrar el método, consideremos un caso con cuatro OTU: *A, B, C y D*. La distancia evolutiva relacional por pares está dada por la siguiente matriz:

Tabla 2.3.1

OTU	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

En esta matriz, d_{ij} equivale a la distancia entre las OTU *i* y *j*. Los dos primeros OTUs en ser utilizados para agrupación son aquellos cuya distancia sea la menor. Supongamos ahora que d_{AB} es la distancia más pequeña. Entonces, los OTU *A* y *B* son los primeros en entrar al grupo y el punto de inserción de la rama, l_{AB}, se posiciona a la distancia de $\frac{d_{AB}}{2}$ sustituciones.

Siguiendo con el proceso de acumulación, *A* y *B* son considerados ahora como un solo OTU, y se calcula una nueva distancia de matrices.

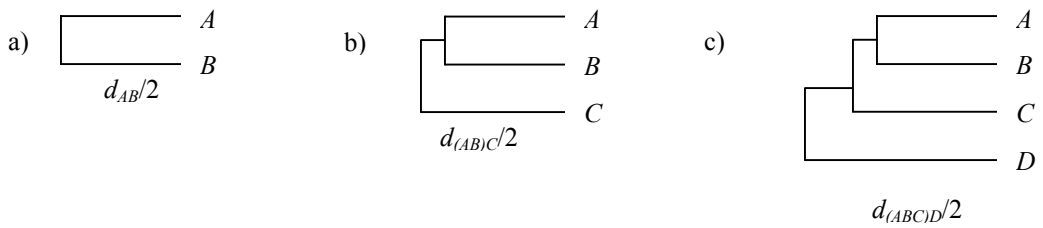
Tabla 2.3.2

OTU	*(AB)*	*C*
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

En esta matriz, $d_{(AB)C} = \frac{d_{AC} + d_{BC}}{2}$ y $d_{(AB)D} = \frac{d_{AD} + d_{BD}}{2}$. Es decir, la distancia entre un OTU simple y un OTU compuesto es el promedio de distancias entre el OTU simple y el constituyente simple de los OTUs compuestos. Si $d_{(AB)C}$ fuera la distancia más corta de la nueva matriz, entonces el OTU *C* se uniría al OTU compuesto *(AB)* con un nodo de distancia $l_{(AB)C} = \frac{d_{(AB)C}}{2}$.

El paso final consiste en realizar la acumulación del último OTU, *D*, con el OTU compuesto, *(ABC)*. La raíz del árbol se coloca a $l_{(ABC)D} = \frac{d_{(ABC)D}}{2} = \frac{\frac{d_{AD} + d_{BD} + d_{CD}}{2}}{3}$. El árbol inferido utilizando UPGMA se muestra a continuación. (Ver Fig. 2.3.3.)

Figura 2.3.3



Descripción del árbol Filogenético en pasos.

En UPGMA, los puntos de ramificación entre dos OTUs simples, *i* y *j*, se colocan a la mitad de la distancia entre ambos.

$$l_{ij} = \frac{d_{ij}}{2}$$

Los puntos de ramificación entre un OTU simple, *i*, y un OTU compuesto, *(jm)*, se coloca al punto medio aritmético de la distancia entre el OTU simples y cada uno de los OTUs constitutivos simples del OTU compuesto.

$$l_{i(jm)} = \frac{\frac{(d_{ij} + d_{im})}{2}}{2}$$

Los puntos de ramificación entre dos OTUs compuestos se posiciona al punto medio aritmético de las distancias entre los constitutivos simples de cada OTU en cada OTU compuesto. Por ejemplo, la posición del punto de ramificación del OTU compuesto *(ij)*, y el OTU compuesto, *(mn)*, es:

$$l_{(ij)(lm)} = \frac{\frac{(d_{im} + d_{in} + d_{jm} + d_{jn})}{4}}{2}$$

En el caso de una OTU compuesta tripartita, *(ijk)*, y una OTU compuesta bipartita, *(mn)*, el punto de ramificación es:

$$l_{(ijk)(lm)} = \frac{\frac{(d_{im} + d_{in} + d_{jm} + d_{jn} + d_{km} + d_{mn})}{6}}{2}$$

UPGMA es uno de los pocos métodos de la reconstrucción filogenética que crean árboles enraizados. Notese además que al utilizar UPGMA se obtiene una topología del árbol y la longitud de las ramas simultáneamente.

La ventaja principal consiste en que con este método se puede tomar en cuenta la mínima evolución que es que cualquier cambio mínimo se ve reflejado en la relación entre las secuencias. Sin embargo, esto se ve reflejado también en una desventaja, ya que una mínima mutación o por cualquier error en la obtención de secuencias, el método tomará en cuenta este error para la construcción del árbol filogenético. [RODIC, 2000]

2.1.3.5 Método de Parsimonia Máxima.

El principio de la parsimonia máxima involucra la identificación de una topología la que requiera el menor número de cambios evolutivos (sustituciones de nucleótidos) para explicar la diferencia observada entre los OTU estudiados. En el método de máxima parsimonia se utilizan estados de caracteres discretos y el camino que lleve al mejor estado que se define por el mínimo número de cambios evolutivos de caracteres, se elige como el mejor árbol. El árbol generado es llamado árbol de máxima parsimonia. Existen muchas ocasiones en el que con el mismo número mínimo de cambios se pueden generar más de un árbol filogenético, de tal manera que no se puede obtener un árbol único. Estos árboles son llamados igualmente parsimoniosos.

Se han desarrollado muchos métodos diferentes para probar diferentes conjuntos de datos. El método que se muestra a continuación fue desarrollado para secuencias de datos de aminoácidos [Eck y Dayhoff, 1966], y modificada posteriormente para secuencias de nucleótidos [Fitch, 1977].

Para empezar, se realiza la clasificación de sitios. El sitio *N* se define como invariante si todos los OTUs bajo estudio poseen el mismo valor en el mismo estado. Los sitios variables pueden ser informativos o no informativos. Un sitio de nucleótidos es filogenéticamente informativo sólo sí favorece a un subgrupo de árboles sobre todos los posibles árboles. Para ilustrar la diferencia entre los sitios informativos y no informativos, consideremos las siguientes secuencias de ADN hipotéticas:

Tabla 2.3.4

	Sitios								
Secuencias	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T
					*		*		*

Existen tres posibles árboles sin raíz para cuatro OTU (tabla 2.3.4). El sitio 1 es no informativo debido a que todas las secuencias en este sitio tienen *A*, así que no es requerido ningún en ninguno de los tres posibles árboles. Para el sitio dos, la secuencia uno tiene *A*, mientras que todos las demás secuencias tienen *G*, y una suposición simple es que el nucleótido *G* cambió a *A* en una guía lineal a la secuencia 1. Por lo cual, es también un sitio no informativo, debido a que de los tres posibles árboles solo se requiere un cambio. Como se muestra en la figura 2.3.4a, para cada posición del árbol del sitio tres requiere dos cambios, y por lo tanto tampoco es informativo. Nótese que si asumimos que el nucleótido que conecta a los OTU 1 y 2 en el árbol I de la figura 2.3.4a es *C* en vez de *G*, el número requerido de cambios para el árbol sigue siendo de dos. La tabla 2.3.4b muestra que para el sitio 4, cada uno de los tres árboles requiere tres cambios, mientras que el árbol II e III requieren dos cambios cada uno (Tabla 2.3.4c). Es por esto que este sitio es informativo. Lo mismo sucede con el sitio 7. El sitio 9 es informativo, pero en contraste con los dos sitios informativos anteriores, éste favorece al árbol II, el cual requiere solo un cambio, mientras que para los árboles I y III se requieren dos cambios para cada uno.

Este ejemplo muestra que los sitios informativos son aquellos que tienen por lo menos dos distintos tipos de nucleótidos en el sitio, cada uno de los cuales se representa en al menos dos de

las secuencias en estudio. En las secuencias anteriores, los sitios informativos (sitio 5, 7 y 9) se indican con asterisco.

Figura 2.3.4

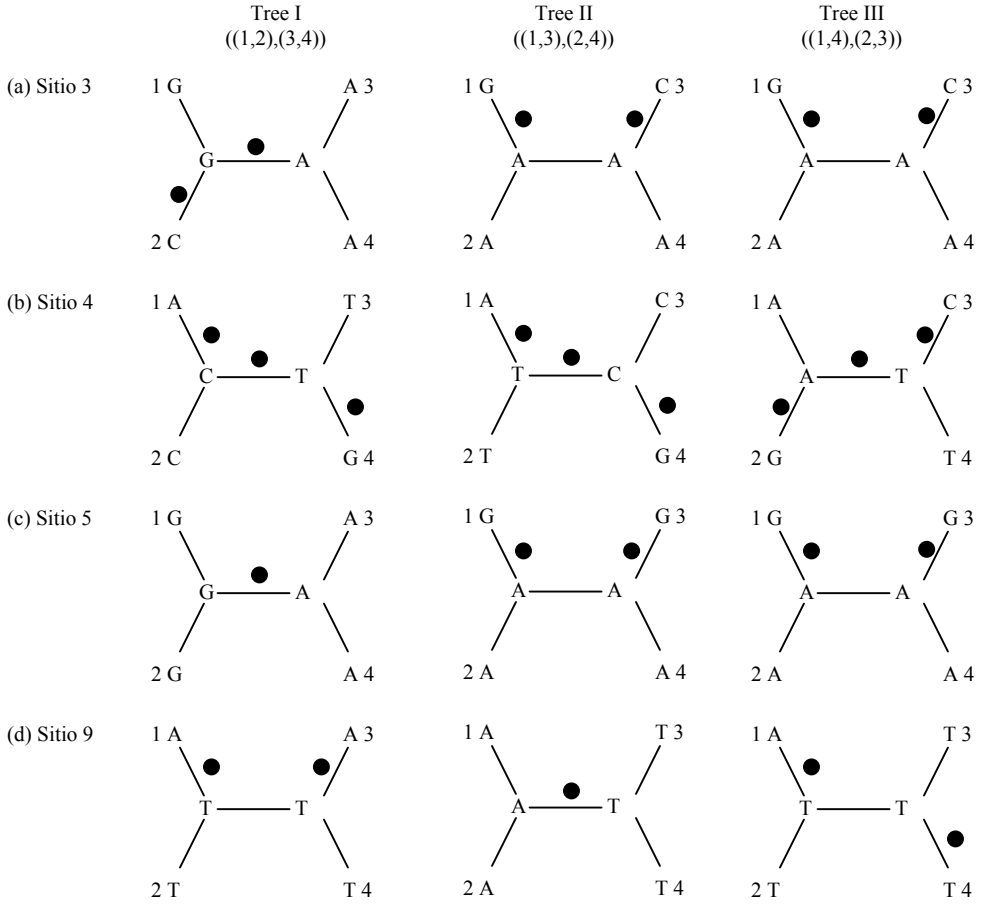


Figura 2.3.4 Tres posibles árboles sin raíz (I, II y III) para cuatro secuencias de DNA (1,2,3 y 4) que se utilizan para escoger el árbol más parsimonioso. La relación ente los posibles árboles de las cuatro secuencias se muestran en el formato Newick. En los nodos terminales se encuentra marcado el número de secuencia y el tipo de nucleótido en la posición homóloga o similar en la especie exenta o que no aplica. Cada punto en la rama significa una inferencia de sustitución en la rama. Noté que el nucleótido a cada lado de los nodos internos del árbol representa una posible reconstrucción de muchas otras alternativas. Por ejemplo, los nucleótidos de ambos lados del nodo interno del árbol III(d) (izquierdo derecho) puede ser una A en vez de la T. En este caso, las dos sustituciones se posicionarán en las ramas que dirige la especie dos y cuatro. Alternativamente, se pueden poner otras combinaciones en los nodos internos. Aunque, esas alternativas requerirían tres o más sustituciones. El número mínino de sustituciones requerido para el sitio nueve es de dos.

Para inferir un árbol de máxima parsimonia se requiere en primera instancia identificar todos los sitios informativos. Posteriormente. se calcula el número mínimo de sustituciones de cada sitio informativo de cada uno de los posibles árboles. En el ejemplo anterior, existen tres sitios informativos. Para los sitios 5, 7 y 9, el árbol I requiere 1, 1 y 2 cambios respectivamente, el árbol II requiere 2, 2 y 1 cambios y el árbol III requiere 2, 2 y 2 cambios. En el último paso, se suma el número de cambios en los sitios informativos para cada posible árbol y se elige el árbol asociado al menor número de cambios. En nuestro caso, el árbol I será seleccionado debido a que

requiere solo 4 cambios en los sitios informativos, mientras que los árboles II y III requieren 5 y 6 cambios, respectivamente.

En el caso de cuatro OTU, un sitio informativo puede favorecer solo a uno de los tres posibles árboles. Por ejemplo, el sitio 5 favorece al árbol I sobre el árbol II y III. En primera instancia, en el ejemplo anterior, el árbol I es el más parsimonioso debido a que está soportado por dos sitios informativos, mientras que el árbol II solo por un sitio y el árbol III por ninguno. En el caso de que existan más de 4 OTU, un sitio informativo puede favorecer a más de un árbol y el árbol más parsimonioso no necesariamente será soportado por el mayor número de sitios informativos.

En el caso de que el número utilizado de OTU sea mayor a cuatro, la situación se vuelve más complicada debido a que existe un número mayor de posibles árboles a considerar.

2.1.3.6 Parsimonia de Fitch-Wagner (Árbol de Wagner).

Bajo el método de máxima parsimonia, se pueden construir diferentes tipos de árboles. Sin embargo, el más aceptado para generar el árbol es el método de Fitch-Wagner en el cual no es requisito conocer el estado ancestral para poder generar el árbol.

Los pasos que se utilizan para generar el árbol de Wagner son los siguientes:

Tomemos un ejemplo que cuente con cinco OTU:

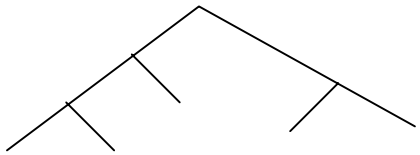
Tabla 2.3.5

X	Total
S1	A
S2	A
S3	G
S4	A
S5	T

El primer paso consiste en seleccionar un árbol *T* cuya longitud sea cero con el número de elementos (OTUs) que se consideren para crear el árbol, la longitud cero se refiere a que la separación entre las ramas es nula teóricamente, ya que no tiene asignado ningún elemento al cual comparar. Como se muestra en la figura 2.3.5

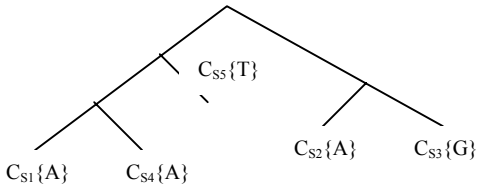
Figura 2.3.5

T =



A continuación se asigna a cada rama un OTU. Como se muestra a continuación:

Figura 2.3.6



Donde C_{sn} es el elemento en la posición *n* de la cadena *C*

Posteriormente se prosigue recorrer el árbol en post-orden. [Graur y Li, 2000].

Algoritmo

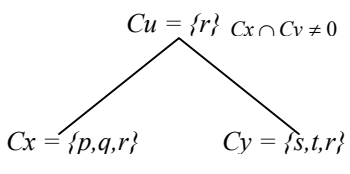
1. Si U es un nodo que cuente con hijos.

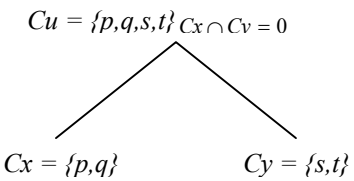
Donde se considera que si $Cx \cap Cy \neq 0$,

1.1.- entonces $Cu = Cx \cap Cy$

1.2.- si no $Cu = Cx \cup Cy$.

Ejemplo:





1. Empezar por la raíz de forma transversal en pre-orden.

2. Tomemos como vértice U el estado de la raíz.

2.1. Si U es raíz entonces $X_{(u,i)}$ *es cualquier elemento de Ci,*

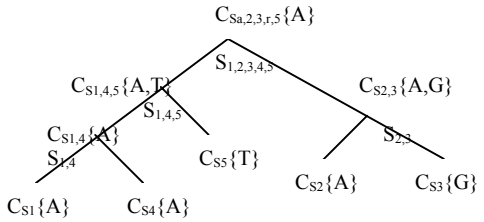
2.2. Si no,

2.2.1. Si U es un vértice interno y tomando 2 como el ancestro de U.

2.2.1.1. Si $X_{(x,i)} \Sigma Cu$ *entonces* $X_{(u,i)} = X_{(x,i)}$.

2.2.1.2. Si no escoger arbitrariamente un elemento de Cu.

3. Por último se obtiene la salida como se muestra:



Este es el método de construir el árbol de Wagner el cual se utiliza en este proyecto.

Las ventajas es que parsimonia aparentemente asume algunos procesos evolutivos y es por esto que todavía es muy controversial su utilización. Para describir este punto se presentan dos agumentos:

El primero se basa en que parsimonia es una convención metodológica que obliga a maximizar el valor de similitud evolutitiva, es decir, obliga a que dos secuencias diferentes en su ADN a que sean similares.

El segundo argumento es que parsimonia asume implícitamente la evolución tomando en cuenta que el cambio evolutivo es escaso. Esto implica que el árbol que minimice los cambios es el mejor estimado. Y desde este punto de vista es similar a un método de máxima similitud.

Viendo estos argumentos podemos ver que el tomar sólo algunas zonas, las cuales se les llama zonas de información, ignore mucha información importante para describir la evolución.

2.2 Algoritmos evolutivos

Las primeras ideas para simular o imitar la evolución natural con el objeto de resolver problemas fueron planteadas por John Von Neumann, incluso antes del descubrimiento del ADN. [MARTÍ,2000] Von Neumann afirmó que la vida debía de estar apoyada por un código que a la vez describiera cómo se puede construir un ser vivo, y tal que ese ser creado fuera capaz de

autorreproducirse; por tanto, un autómata o máquina autorreproductiva tendría que ser capaz, además de contener las instrucciones para hacerlo, de copiar tales instrucciones a su descendencia.

Sin embargo, no fue hasta mediados de los años cincuenta, cuando el rompecabezas de la evolución fue prácticamente completado, cuando Box comenzó a pensar en imitarla para, en su caso, mejorar procesos industriales. La técnica de Box, denominada EVOP (Evolutionary Operation), consistía en elegir una serie de variables que regían un proceso industrial. Sobre esas variables se creaban pequeñas variaciones que formaban un hipercubo, variando el valor de las variables una cantidad fija. Se probaba entonces con cada una de las esquinas del hipercubo durante un tiempo, y al final del periodo de pruebas, un comité humano decidía sobre la calidad del resultado. Es decir, se estaba aplicando mutación y selección a los valores de las variables, con el objeto de mejorar la calidad del proceso. Este procedimiento se aplicó con éxito a algunas industrias químicas. [MARTÍ,2000]

Un poco más adelante, en 1958, Friedberg y sus colaboradores pensaron en mejorar el funcionamiento de un programa usando técnicas evolutivas. Para ello diseñaron un código máquina de 14 bits, y cada programa tenía 64 instrucciones. Un programa llamado Herman, ejecutaba los programas creados, y otro programa, el Teacher o profesor, le mandaba a Herman ejecutar otros programas y ver si los programas ejecutados habían realizado su tarea o no. La tarea consistía en leer unas entradas, situadas en una posición de memoria, y debían depositar el resultado en otra posición de memoria, que era examinada al terminarse de ejecutar la última instrucción. [MARTÍ,2000]

Para hacer evolucionar los programas, Friedberg hizo que en cada posición de memoria hubiera dos alternativas; para cambiar un programa, alternaba las dos instrucciones (que eran una especie de alelos), o bien reemplazaba una de las dos instrucciones con una totalmente aleatoria.

En realidad, lo que estaba haciendo es usar mutación para generar nuevos programas; al parecer, no tuvo más éxito que si hubiera buscado aleatoriamente un programa que hiciera la misma tarea. El problema es que la mutación sola, sin ayuda de la selección, hace que la búsqueda degenere en una búsqueda aleatoria.

Más o menos simultáneamente, Bremmerman trató de usar la evolución para ``*entender los procesos de pensamiento creativo y aprendizaje''*, y empezó a considerar la evolución como un proceso de aprendizaje. Para resolver un problema, codificaba las variables del problema en una cadena binaria de 0s y 1s, y sometía la cadena a mutación, cambiando un bit cada vez. Bremmerman trató de resolver problemas de minimización de funciones, aunque no está muy claro qué tipo de selección usó y el tamaño y tipo de la población. En todo caso, se llegaba a un punto, la ``trampa de Bremmerman'', en el cual la solución no mejoraba; en intentos sucesivos trató de añadir entrecruzamiento entre soluciones, pero tampoco obtuvo buenos resultados. Una vez más, el simple uso de operadores que creen diversidad no es suficiente para dirigir la búsqueda genética hacia la solución correcta; y esto se logra aplicándolo a un concepto de la evolución darwiniano clásico: por mutación, se puede mejorar a un individuo; en realidad, la evolución actúa a nivel de población. [AGUADÉ,2000]

El primer uso de procedimientos evolutivos en computación se debe a Reed, Toombs y Baricelli, que trataron de hacer evolucionar un tahúr que jugaba a un juego de cartas simplificado. Las estrategias de juego consistían en una serie de 4 probabilidades de apuesta alta o baja con una mano alta o baja, con cuatro parámetros de mutación asociados. Se mantenía una población de 50 individuos, y aparte de la mutación, había intercambio de probabilidades entre dos padres. Es de suponer que los perdedores se eliminaban de la población (tirándolos por la borda). Aparte de, probablemente, crear buenas estrategias, llegaron a la conclusión de que el entrecruzamiento no aportaba mucho a la búsqueda. [MARTÍ,2000]

2.2.1. Algoritmos Genéticos

Los Algoritmos Genéticos (GA) fueron introducidos por John Holland en 1970 siendo inspirado en el proceso observado en la evolución natural de los seres vivos. [AGUADÉ,2000]

Los biólogos han estudiado profundamente los mecanismos de la evolución, y aunque quedan parcelas por entender, muchos aspectos están bastante explicados. De manera muy general, podemos decir que en la evolución de los seres vivos el problema al que cada individuo se enfrenta cada día es la supervivencia. Para ello cuenta con las habilidades innatas provistas en

su material genético. A nivel de los genes, el problema es buscar aquellas adaptaciones benéficas en un medio hostil y cambiante. Debido en parte a la selección natural, cada especie gana cierta cantidad de "conocimiento", el cual es incorporado a la información de sus cromosomas. [AGUADÉ,2000]

Así pues, la evolución tiene lugar en los cromosomas, en donde está codificada la información genética del ser vivo. La información almacenada en el cromosoma varía de una generaciones a otra. En el proceso de formación de un nuevo individuo, se combina la información de los cromosomas de los progenitores aunque la forma exacta en que se realiza es aún desconocida. [MARTÍ,2000]

Aunque muchos aspectos están todavía por discernir, existen unos principios generales ampliamente aceptados por la comunidad científica. Algunos de éstos son:

La evolución opera en los cromosomas más no en los individuos a los que representan.

La selección natural es el proceso por medio del cual los cromosomas con "buenas estructuras" se reproducen más a menudo que los demás.

En el proceso de reproducción tiene lugar la evolución mediante la combinación de los cromosomas de los progenitores. Llamamos recombinación al proceso en el que se forma el cromosoma del descendiente. Este proceso de recombinación está sujeto a las mutaciones que pueden alterar dichos códigos.

La evolución biológica no tiene memoria en el sentido de que en la formación de los cromosomas de un descendiente únicamente se considera la información de la generación anterior.

Los algoritmos genéticos establecen una analogía entre el conjunto de soluciones de un problema y el conjunto de individuos de una población codificando la información de cada solución en una secuencia (vector binario) a modo de cromosoma. En palabras del propio Holland: "Se pueden encontrar soluciones aproximadas a problemas de gran complejidad computacional mediante un proceso de "evolución simulada". [MARTÍ,2000]

A tal efecto se introduce una función de evaluación de los cromosomas, que llamaremos aptitud y que está basada en la función objetivo del problema. De la misma manera, se introduce un mecanismo de selección de manera que los cromosomas con mejor evaluación sean escogidos para "reproducirse" más a menudo que los demás.

Los algoritmos desarrollados por Holland eran conceptualmente simples, pero demostraron la capacidad de producir buenos resultados en problemas complejos en diversias áreas. Los algoritmos Genéticos están basados en integrar e implementar eficientemente dos ideas fundamentales: Las representaciones simples como secuencias binarias de las soluciones del problema y la realización de transformaciones simples para modificar y mejorar estas representaciones.

Para llevar a la práctica el esquema anterior y concretarlo en un algoritmo, hay que especificar los siguientes elementos:

· Una representación de los individuos
· Una población inicial
· Una medida de evaluación
· Un criterio de selección / eliminación de cromosomas
· Una o varias operaciones de recombinación
· Una o varias operaciones de mutación

En la actualidad podemos distinguir dos enfoques predominantes: Limitarse a cadenas binarias o utilizar otro tipo de configuraciones. Hemos de notar que las operaciones genéticas dependen del tipo de representación, por lo que la elección de una condiciona a la otra.

La ventaja de las primeras es que permite definir fácilmente operaciones de recombinación, además los resultados sobre convergencia están probados para el caso de cadenas binarias. Sin embargo, el utilizarlas en ciertos problemas puede ser poco natural y poco eficiente. Por ejemplo, en el problema del agente viajero sobre 5 ciudades y 20 aristas, la cadena 01000100001000100010 representa una solución sobre las aristas ordenadas. Sin embargo, dicha representación no es muy natural y además, no todas las cadenas con cinco unos representan soluciones lo cual complica substancialmente la definición de una operación de recombinación.

Es más natural la ruta de ciudades: (2,3,1,5,4), lo cual permite definir naturalmente diferentes operaciones estables. [AGUADÉ,2000]

La población inicial suele ser generada aleatoriamente. Sin embargo, recientemente se están utilizando métodos heurísticos para generar soluciones iniciales de buena calidad. En este caso, es importante garantizar la diversidad estructural de estas soluciones para tener una "representación" de la mayor parte de población posible o al menos evitar la convergencia prematura que es que la población converja a un mínimo o máximo local en vez de tender al global. [MARTI,2000]

Respecto a la evaluación de los cromosomas, se suele utilizar la calidad como medida de la bondad según el valor de la función objetivo en el que se puede añadir un factor de penalización para controlar la infactibilidad. Este factor puede ser estático o ajustarse dinámicamente, lo cual produciría un efecto similar al de la Oscilación Estratégica en Tabu Search: [MARTÍ,2000]

CALIDAD = VALOR OBJETIVO NORMALIZADO – PENALIZACIÓN x MEDIA INFACTIBILIDAD.

Más o menos a mediados de los años 60, Rechenberg y Schwefel describieron las estrategias evolutivas (EE). Las estrategias evolutivas son métodos paramétricos de optimización, que trabajan con poblaciones de cromosomas compuestos por números reales. Hay diversos tipos de estrategias de evolución, que se verán más adelante. En la más común, se crean nuevos individuos de la población añadiendo un vector mutación a los cromosomas existentes en la población; en cada generación, se elimina un porcentaje de la población, y los restantes generan la población total, mediante mutación y cruce. La magnitud del vector mutación se calcula adaptativamente. Una revisión sobre las estrategias evolutivas se puede encontrar en Bäck, Hoffmeister y Schwefel. [AGUADÉ,2000]

A partir de los años 60 se han desarrollado algoritmos o métodos que podríamos llamar evolutivos modernos, y se han seguido investigando hasta nuestros días. Algunos de ellos son simultáneos a los algoritmos genéticos, pero se desarrollaron independientemente sin conocimiento unos de otros. Uno de ellos, la programación evolutiva (PE) de Fogel, Owens y Walsh, se inició como un intento de usar la evolución para crear máquinas inteligentes, que pudieran prever su entorno y reaccionar adecuadamente a él. Para simular una máquina pensante,

se utilizó un autómata finito. Un autómata finito es un conjunto de estados y reglas de transición entre ellos, de forma que, al recibir una entrada, cambia o no de estado y produce una salida. [MARTÍ,2000]

Fogel trataba de hacer aprender a estos autómatas a encontrar regularidades en los símbolos que se le iban enviando. Como método de aprendizaje usó un algoritmo evolutivo: una población de diferentes autómatas competía para hallar la mejor solución, es decir, predecir cual iba a ser el siguiente símbolo de la secuencia con un mínimo de errores; los peores 50% eran eliminados cada generación, y sustituidos por otros autómatas resultantes de una mutación de los existentes.

De esta forma, se lograron hacer evolucionar autómatas que predecían algunos números primos (por ejemplo, uno de ellos, cuando se le daban los números más altos, respondía siempre que no era primo; la mayoría de los números mayores de 100 son no primos). En cualquier caso, estos primeros experimentos demostraron el potencial de la evolución como método de búsqueda de soluciones novedosas.

2.2.2. Programación Genética

La programación genética (PG) se basa en la evolución de programas. Normalmente la representación usada sigue la sintaxis de programas en LISP, que esencialmente son árboles cuyos nodos internos son operadores y los externos operandos. Este tipo de estructura de datos requiere el uso de operadores de mutación y cruce específicos. [AGUADÉ,2000]

Michalewicz propuso en su libro ``*Genetic Algorithms + Data Structures = Evolution Programs''* prescindir de la representación usual de los individuos en la población (en cadenas de bits o vectores de números reales, como se venía haciendo), y al mismo tiempo aplicar el paradigma principal de la programación procedural a la computación evolutiva: aplicar algoritmos a estructuras de datos, ya que son distintas y deben estar separadas, y formar programas evolutivos mediante la unión e interacción de ambos.

2.2.3. Construcción de filogenias utilizando algoritmos genéticos.

Existen varios sistemas que utilizan heurísticas para generar árboles filogenéticos, pero el más representativo, fue creado por Clare Bates Cogdon, quien propuso una aplicación basada en algoritmos genéticos para construcción de filogenias, que es un acercamiento utilizado por los biólogos para estudiar la relación evolutiva entre los organismo y lo llamó "Gaphyl: An Evolutionary Algorithms Approach for the Study of Natural Evolution" [BATES,2000],. La diferencia entre los métodos de construcción de filogenias anteriores y Gaphyl radica en que a diferencia de los métodos tradicionales de construcción de filogenias que buscan de manera exhaustiva el mejor modelo evolutivo, Gaphyl lo realiza por medio de búsquedas heurísticas para buscar la mejor hipótesis de evolución, debido a que con el método de búsqueda exhaustiva el problema se vuelve poco práctico. Es de esta manera que Gaphyl puede producir buenos resultados en mucho menor tiempo que por métodos tradicionales. Gaphyl se basó en dos paquetes computacionales utilizados ampliamente que son "PHYLIP" y "Genesis", el primer paquete es un sistema que agrupa varias herramientas para la evaluación de filogenias, mientras que el segundo es un sistema que apoya a la realización de experimientos utilizando algoritmos genéticos, Bates fusionó ambos métodos utilizando algoritmos genéticos para generar los árboles y la evaluación de Phylip del paquete de Parsimonia para evaluar dichos árboles.

Gaphyl utilizó datos Lamiiflorae y Angiosperma, mutación de intercambio de posición entre dos árboles y cruza de padres (árboles) buscando las ramas en donde se encontraban las hojas de un padre en el otro e intercambiando posiciones. De esta manera Gaphyl comprobó que podía generar más árboles de una aptitud apta de parsimonia en el mismo tiempo que con Phylip.

Otros investigadores han utilzado algoritmos genéticos para la construcción de filogenias como "Using genetic algorithms for the construction of phylogenetic trees: application to G-protein coupled receptor sequences [Reijmers et al,1998]. Quienes proponen métodos utilizando algoritmos genéticos para construcción de filogenias que tiendan a buenos árboles, sin depender de la población inicial. Ellos comprobarón con Phylip la funcionalidad de su método.

A pesar de que existen muchos métodos para la generración de filogenías en base a Algoritmos Evolutivos y de que la mayoría son muy aceptados por su funcionalidad y rapidez, no han considerado un aspecto muy importante. Este aspecto es la función objetivo. Parsimonia solo hace agrupaciones de sitios informativos (clusters) ignorando toda la mutación excedente. De esta manera se podrán producir varios árboles sin embargo no siempre serán los mejores.

3. Modelo Propuesto.

El modelo propuesto en esta investigación es un método de generación de árboles filogenéticos basados en algoritmos genéticos con una función objetivo híbrida debido a que creemos que la unión de dos métodos de evaluación proporcionan un enfoque más completo.

La hipótesis de esta investigación establece que el uso de los algoritmos evolutivos tales como los algoritmos genéticos, basados en más de un criterio de evaluación para calcular la aptitud en la construcción de filogenias produce un árbol filogenético que compensa las debilidades de una metodología con las fortalezas del otro método logrando balancear la estructura y la similitud entre los elementos comparados y clasificar más eficiente la relación entre las especies además de mejorar la velocidad de procesamiento.

Para poder definir la relación que existe entre las especies o cadenas a comparar, la mutación juega un papel importante. Es sabido que alrededor del 2.4% de la información genética tiene una mutación a nivel genético. Pero esta mutación se puede originar no solo cuando se genera un nuevo ser, sino también cuando el ADN se transcribe o el ARN se traduce. Así mismo el ADN sufre trastornos a lo largo de la vida de un ser, siendo este también una mutación. [KIMURA,1983]

Teniendo en cuenta todos estos aspectos de la mutación, es imposible ignorar completamente la mutación en todos los aminoácidos o proteínas, pero también no es certero tomar en cuenta la mutación en su totalidad, debido a que puede ser solo un cambio local que no tiene mucho significado y que no permitirá fijar una relación entre las especies.

Para poder equilibrar esta evaluación y poder tomar en cuenta la mutación y también tomar el cuenta el patrón de similitud y minimizar el efecto de la mutación se propone hibridizar la función objetivo para la generación de árboles filogenéticos.

El valor de normalización para decidir cuanto tomar en cuenta el valor de la mutación, se utiliza la teoría neutralista de la evolución molecular. Fijando así al valor de normalización para asignarle a DMM un valor de 2.4 % que es el que se establece como valor promedio de evolución molecular para los Aminoácidos y 0.8 % para las proteínas. [KIMURA,1983]

Estos valores se basan en estudios realizados por Motoo Kimura, y en su opinión, la mayoría de los genes mutantes son selectivamente neutros, es decir, no tienen selectivamente ni más ni menos ventaja que los genes a los que sustituyen; en el nivel molecular, la mayoría de los cambios evolutivos se deben a la deriva genética de genes mutantes selectivamente equivalentes.

El elevado número de combinaciones requeridas para poder encontrar la mejor solución es un gran problema, esto se debe a que el espacio de búsqueda es muy grande y de ahí que resulte necesaria la utilización de heurísticas que permitan llegar a una solución aceptable, sin tener que realizar la búsqueda en su totalidad.

Para resolver este problema, se utiliza el algoritmo genético, el cual es una herramienta utilizada para realizar búsquedas heurísticas de manera efectiva, ya que se sabe que el algoritmo genético produce resultados muy certeros simulando el modelo evolutivo Darwiniano sobre evolución y utiliza la mutación, reproducción y selección por médio de una función de aptitud o función objetivo, esta función objetivo es la que se modifica para hibridizar y tomar en cuenta la aptitud del individuo que evaluamos. La función objetivo permite la hibridación debido a las ecuaciones que se toman como función de evaluación tienen la capacidad de ser simples o compuestas es decir de un solo objetivo o multiobjetivo, así mismo se pude tener una ecuación que junte dos ecuaciones a lo cual llamamos híbrido. [MARTÍ,2000]

Clare Bates Cogdon en su investigación “Gaphyl: An Evolutionary Algoritms Approach for the Study of Natural Evolution” [BATES,2000], utiliza algorimos genéticos con máxima parsimonia para generar los árboles filogenéticos, de esta manera se comprueba la efectividad de los algoritmos genéticos en cuanto a la velocidad y funcionalidad para obtener árboles válidos en un tiempo razonable.

3.1. Método híbrido de Evaluación de árboles filogenéticos.

Después de mencionar la hipótesis, podemos ver que lo que se busca es crear un método de evaluación el cual no considera la totalidad de la mutación pero que tampoco la ignore por completo.

El método de generación de árboles filogenéticos por medio de parsimonia máxima busca sitios específicos donde se pueda crear una agrupación (cluster) y de esta manera clasificar en esos sitios (sitios informativos) las especies con el menor número de cambios evolutivos para relacionarlos.

Por otra parte, el método de matríz de distancia es su modelo de máxima similitud. La función que realiza es la de comparar dos especies en cuanto a la diferencia entre las secuencias y compararlas sucesivamente con las demás para encontrar el menor número, es decir, aquellas que sean las más parecidas.

Lo que notamos en estos dos casos es que máxima parsimonia, a pesar de ser uno de los métodos con mayor aceptación, ignora la mayoría de las secuencias en las cadenas de las especies sino que encuentra un punto en el que pueda lograr una agrupación, lo cual implica que se está ignorando una cantidad de información importante para poder clasificar de manera precisa ya que omite mucha mutación. Pero en su contraparte existe el método de máxima similitud, el cual compara cada elemento de la cadena de la especie, por lo que se está tomando en cuenta hasta la mutación mínima de la cual podría ser derivada de un error en la obtención de la información, en la traducción o transcripción de la cadena o simplemente una mutación que nunca se fijará en la especie.

El método híbrido fusiona ambas formas de generación de árboles filogenéticos: el método de parsimonia máxima y la de máxima similitud, con el propósito de evitar ignorar la mayoría de la mutación y buscar agrupaciones, así como evitar el de tomar en cuenta la totalidad de la mutación que muchas veces sólo es un cambio irrelevante.

Para poder fusionar los dos métodos se utiliza la función objetivo del algoritmo genético, ya que por su naturaleza, al igual que en los algoritmos genéticos con función objetivo multiple o multiobjetivo, el modelo se presta para poder realizar varias evaluaciones en una sola agrupación de ecuaciones. Esta medida, además del volumen de combinaciones necesarias para la búsqueda, crean la necesidad considerar técnicas heurísticas. En este contexto, el algoritmo genético se propone en este trabajo como una alternativa para la generación del modelo.

Para crear el árbol es necesario que las secuencias se encuentren alineadas en pares, esta alineación implica que el conjunto de secuencias a comparar tienen que poder ser comparables en sus bases y para esto se les inserta espacios logrando así hacer que concuerden sus bases por ejemplo:

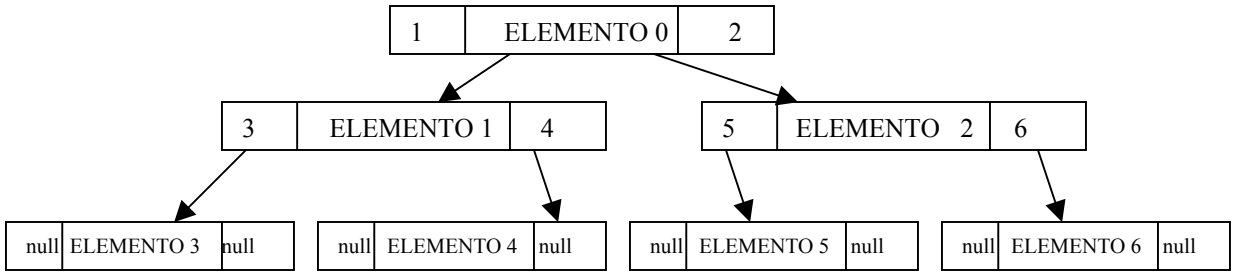
Figura 3.1.1 Alineación se Secuencias

CCTT**CA**GAATA**CAG**AATAGG**G**ACATA**G**AGA
ATCC**CA**_ _ _CC**CAG**CCCCCT**G**GACCT**G**TAT

Posteriormente, el sistema creará varios árboles para evaluarlos con la función objetivo, de ellos el mejor pasará a la siguiente generación automáticamente para prevenir eliminar el mejor elemento. El siguiente paso es la selección, cruza y mutación cuya ocurrencia está sujeta a cierta probabilidad.

Para la primera generación, los individuos de la población serán creados de manera aleatoria para producir así diferentes árboles. Un árbol se representa como un conjunto de arreglos ordenados en su posición, es decir, en un método de representación del modelo computacional de la relación de los elementos o especies con respecto a otras y se muestra de la siguiente forma:

Figura 3.1.2 Representación de Árboles por Listas



En el diagrama anterior se muestra que el elemento 0 está ligado al elemento 1 que es menor y al elemento 2 que es mayor. Tanto el elemento 1 como el 2 no son nodos terminales. El número uno del elemento 0 significa la posición en el arreglo de la misma forma el número 2, la única diferencia es si es menor o mayor al elemento 0. Y sucesivamente hasta llegar a los elementos 3, 4, 5 y 6.

Para el experimento se generaron dos tipos de programas, uno que lee ADN y el otro que lee proteínas. En el primer experimento se utilizaron secuencias de ADN de virus de inmunodeficiencia humana, sin embargo, el problema es que el ADN codifica proteínas en tripletes conocidos como codones, al existir diferentes codones que son conformados por los cuatro tipos de aminoácidos existentes, por lo que se generan $4 \times 4 \times 4 = 64$ combinaciones y existen solo 20 proteínas, por lo que más de un codón produce la misma proteína. Debido a esto se tomó en cuenta el segundo experimento que utiliza proteínas, esto también a que expertos del IBT quienes validaron la salida, conocían de manera profunda la relación del segundo experimento.

La razón por la que se utilizó el VIH se debe a que existen experimentos tales como la búsqueda de epidemiología en la transmisión de enferemedades de prestadores de servicio de salud a sus pacientes explicado por Ou, et. al., en donde se describe un estudio realizado en pacientes contagiados por un dentista en Miami. El método que se propone ayuda a poder realizar estudios de epidemiología por lo cual se utilizarón las secuencias de ADN de HIV.[OU,1992]

Los datos de entrada se muestran en el ANEXO 1 donde se ven las cadenas utilizadas.

3.2. Diseño del algoritmo genético.

Los métodos heurísticos de búsqueda han sido ampliamente utilizados para la generación de árboles filogenéticos, esto se debe a que el volumen de información a procesar para la búsqueda del mejor árbol es muy grande y utilizar métodos de evaluación directos pueden ser muy costosos computacionalmente hablando. Existen diferentes tipos de heurísticas tales como el recocido simulado y los algoritmos genéticos los cuales han sido los que en su mayoría se utilizan como en el caso de Gaphyl.

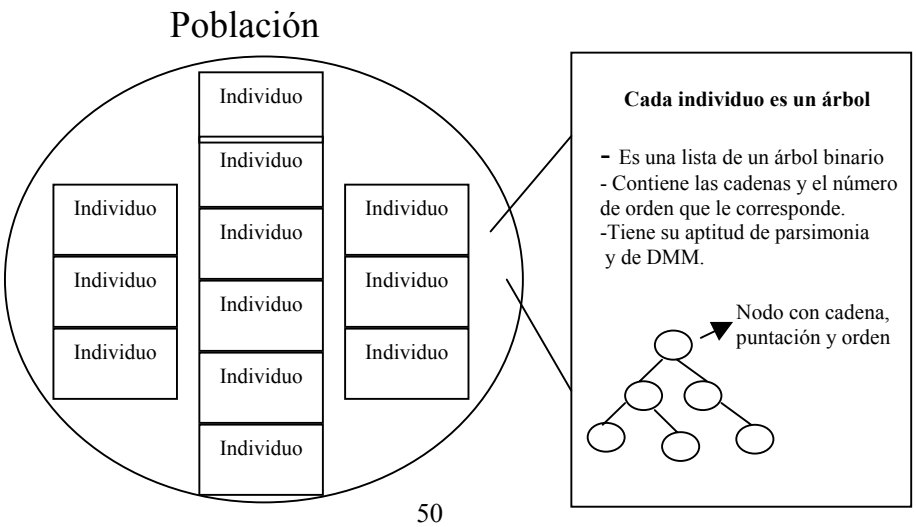
El sistema utiliza el algoritmo genético como base heurística ya que el volumen de información a procesar es muy grande para resolver problemas de generación de árboles filogenético. Asimismo, la función objetivo se presta para poder tener más de un valor el cual sirva para evaluar al mejor árbol.

La estructura del la población del algoritmo genético es de la siguiente manera:

Cada individuo de una población es un árbol que tiene nodos cada uno con su cadena (ADN o Proteina), su aptitud y el orden de los nodos que constituyen el árbol.

En diagrama sería de la siguiente forma:

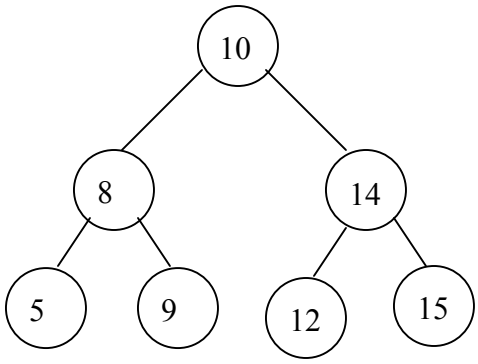
Figura 3.2.1 Población



La estructura del árbol como elemento de la población se describe como cada árbol tiene sus nodos la comparación se realiza comparando el hijo izquierdo del nodo con el padre y el hijo derecho con el izquierdo y bajar al siguiente nivel. Por ejemplo:

Supongamos un arreglo que representa un árbol binario por ejemplo:

Figura 3.2.2 Árbol Binario



Posición	1	2	3	4	5	6	7
Valor	10	8	14	5	9	12	15
Cadena	AGC	TCA	ATT	TAG	TGA	AAA	AGT

El algoritmo para el acomodo es el siguiente:

- *El nodo raíz es la posición uno.*
- *El hijo izquierdo está dado por el doble de la posición del nodo, es decir si la posición es X entonces el hijo izquierdo será 2X .*
- *El hijo derecho está dado está dado por el doble de la posición del nodo más uno, es decir si la posición es X entonces el hijo derecho será 2X + 1.*

En el ejemplo anterior, podemos observar que el hijo izquierdo del nodo raíz con valor de 10 y que ocupa la posición 1 es *2X1 = 2*. La posición dos tiene valor de 8 que como se ve en la figura es el hijo izquierdo y el hijo derecho de la raíz entonces sería *(2X1) + 1 = 3* es decir el número 14. Y de manera sucesiva se describe el árbol binario.

Entonces la población es un arreglo de árboles:

Tabla 3.2.1 Evaluación de árboles

Árbol	1	2	3	4	5	6	7
Puntuación	10	8	14	5	9	12	15

Para iniciar el algoritmo, el sistema genera árboles de manera aleatoria ubicando a cada nodo una especie y validando que no se repita la especie en un mismo árbol, de esta manera se generan 10 árboles los cuales son diferentes y conforman la población, se escogió una población de 10 ya que se probarón con más árboles y el tiempo de convergencia es el mismo debido a que el mejor siempre se conserva en las siguientes generaciones (elitismo), además de que el mejor tiene una probabilidad grande de cruce. Asímismo, se le asigna a cada rama o nodo, la secuencia del árbol que es el ADN o Proteina que debe de estar alineado y la longitud es variable así como el número de ramas.

Una vez inicializando el algoritmo, el siguiente paso es realizar la búsqueda heuristica de la evolución realizando la evaluación de aptitud de los árboles, cruza y mutación. Este ciclo se realizará hasta que se cumpla alguna de las dos condiciones de paro.

La primera es que el error cuadrático medio, el cual es la estabilidad bajo error cuadrático medio establecido en cuanto a que la diferencia cuadrática sea menor a una centésima, esto quiere decir que si no exista cambios en la población en un número determinado de iteraciones. Se escoge una centésima ya que de esta manera el error es lo suficientemente pequeño, es decir, la diferencia entre las generaciones es mínima aunque todavía existe pero tiene una precisión de $\frac{1}{100}$. De una manera más matemática, la ecuación sería:

$$\frac{(ppant - ppact)^2}{2} = ecm$$

Donde:
ppant es la puntuación promedio de la población anterior.
pact es la puntuación promedio de la población actual.
ecm es el error cuadrático médio.

Y el segundo criterio de paro es el número de iteraciones que en el caso de este proyecto fue de mil iteraciones.

3.2.1 La función objetivo.

La función objetivo, que es la que define la aptitud y asigna la puntuación por lo que está ligado al error cuadrático médio, se divide en dos partes, la primera evaluación utiliza UPGMA (*Distance Matriz Method*) y la segunda con Método de Parsimonia Máxima.

En el primer caso (UPGMA), para cada árbol en la población, se evaluará utilizando OTU's en la forma $D(a,b,...,n)$ donde n es el número de elementos del árbol. La evaluación de D se refiere al número de elementos diferentes en ambas cadenas exceptuando los saltos. La ecuación utilizada es $D_{i,j} = \frac{D_i + D_j}{2}$ Esta ecuación será aplicada hasta llegar a "n" para obtener la evaluación.

En el experimento, se toma el elemento del arreglo de la siguiente forma:

Figura 3.2.3 Algoritmo de comparación de secuencias para DMM.

- *Tomar el nodo padre y comparar con el nodo hijo izquierdo, para cada proteína o aminoácido diferente, se aumenta el valor es decir se cuentan las diferencias en los aminoácidos o proteínas y se obtiene el primer valor.*
- *Tomar el nodo padre y comparar con el nodo hijo derecho, para cada proteína o aminoácido diferente, se aumenta el valor es decir se cuentan las diferencias en los aminoácidos o proteínas y se obtiene el siguiente valor.*
- *Se suman los dos valores.*
- *Se toma al nodo hijo como padre y se repite el proceso.*
- *En caso de que no exista hijo termina ese proceso hasta que ninguno tenga hijos.*

Utilizando el ejemplo de la descripción de la estructura tenemos:

Tabla 3.2.2 Evaluación de secuencias por DMM

Posición	1	2	3	4	5	6	7
Valor	10	8	14	5	9	12	15
Cadena	AGC	TCA	ATT	TAG	TGA	AAA	AGT

Figura 3.2.4 Algoritmo de comparación de secuencias para máxima parsimonia.

- Comparar 1 y 2, AGC y TCA existen 3 diferencias.
- Comparar 1 y 3, AGC y ATT existen 2 diferencias.
- Diferencia hasta este punto es de 5.
- Se toma al nodo 2 como padre.
- Comparar 2 y 4, TCA y TAG existen 2 diferencias.
- Comparar 2 y 5, TCA y TGA existen 1 diferencias.
- Diferencia hasta este punto es de 8.
- Se toma el nodo 3 como padre.
- Comparar 3 y 6, ATT y AAA existen 2 diferencias.
- Comparar 3 y 7, ATT y AGT existen 2 diferencias.
- Diferencia hasta este punto es de 12.
- Se toma el nodo 4 como padre.
- No existen más nodos, termina.
- La puntuación final es de 12.

Es así como se obtiene la puntación de DMM en el sistema para posteriormente pasar a máxima parsimonia.

En el segundo caso (Máxima Parsimonia), primero, el algoritmo localiza sitios informativos y no informativos. Posteriormente, cada árbol se evaluará comparando los sitios informativos de cada rama del árbol contra la otra rama más cercana para obtener el valor por parsimonia del árbol.

En el experimento, se toma el elemento del arreglo de la siguiente forma:

Figura 3.2.5 Algoritmo de comparación de secuencias.

- Comparar todas las ramas del árbol analizado para buscar patrones, estos patrones serán aquellos que en la cadena de ADN o proteina, se pueda agrupar, es decir, que aproximadamente la mitad tenga un aminoácido o una proteína y la otra mitad una diferente pero igual entre ellas. Así se encuentra lo que son los sitios informativos.
- Tomar el nodo padre y comparar con el nodo hijo izquierdo, para cada proteína o aminoácido diferente, se aumenta el valor es decir se cuentan las diferencias en los aminoácidos o proteínas de los sitios informativos y se obtiene el siguiente valor.
- Tomar el nodo padre y comparar con el nodo hijo derecho, para cada proteína o aminoácido diferente, se aumenta el valor es decir se cuentan las diferencias en los aminoácidos o proteínas de los sitios informativos y se obtiene el siguiente valor.
- Se suman los dos valores.

- Se toma al nodo hijo como padre y se repite el proceso.
- En caso de que no exista hijo termina ese proceso hasta que ninguno tenga hijos.

Utilizando el ejemplo de la descripción de la estructura tenemos:

Tabla 3.2.3 Evaluaciones de máxima parsimonia.

Posición	1	2	3	4	5	6	7
Valor	10	8	14	5	9	12	15
Cadena	AGC	TCA	ATT	TAG	TGA	AAA	AGT

Figura 3.2.6 Algoritmo con la tabla 3.2.3.

- Identificar sitios informativos, (aquellos que se puedan clasificar, es decir que cada grupo esté formado de un aminoácido y la otra parte de otro).
Solo existe un sitio informativo que es el 1 ya que 4 elementos tienen A (Adenina) y los otros 3 elementos tienen T (Timina) por lo que el sitio informativo es 1.
- Comparar 1 y 2, AGC y TCA en el sitio 1 existe 1 diferencia.
- Comparar 1 y 3, AGC y ATT en el sitio 1 existen 0 diferencias.
- Diferencia hasta este punto es de 1.
- Se toma al nodo 2 como padre.
- Comparar 2 y 4, TCA y TAG en el sitio 1 existen 0 diferencias.
- Comparar 2 y 5, TCA y TGA en el sitio 1 existen 0 diferencias.
- Diferencia hasta este punto es de 1.
- Se toma el nodo 3 como padre.
- Comparar 3 y 6, ATT y AAA existen 0 diferencias.
- Comparar 3 y 7, ATT y AGT existen 0 diferencias.
- Diferencia hasta este punto es de 1.
- Se toma el nodo 4 como padre.
- No existen más nodos, termina.
- La puntuación final es de 1.

Una vez obtenida la evaluación para cada elemento de la población (árbol), se procede a evaluar la función híbrida final en la forma: $(UPMGA \times 0.024) + MAX.PARSIMONIA = EVALUACIÓN.$

Para explicar la función objetivo híbrido, es importante tomar en cuenta el valor por el que se suma o multiplica cada valor. En la teoría neutralista que se explica en el capítulo 2 establece que solo el 2.4 % de las mutaciones tiene una fijación la cual sirva como una evolución.

El método de UPMGA cuenta todas las mutaciones de todas las cadenas de las especies comparadas por lo que está tomando en cuenta todo tipo de mutación, esto se puede ver debido a que compara al nodo padre con sus hijos en todos los aminoácidos o proteínas de la cadena en busca de cualquier cambio. Por otra parte, máxima parsimonia ignora la mayor parte de la mutación y solo toma en cuenta pequeñas zonas donde puede realizar una clasificación. Por esta razón se multiplica a UPMGA por el 2.4% para lograr tomar encuenta la mutación que se fija y así mismo poder realizar una agrupación correcta normalizando el valor de UPMGA con el de máxima parsimonia.

En el ejemplo que tenemos sería:

$$.024 \; X \; 12 + 1 = 1.288$$

Como podemos observar, el valor de UPMGA que era mucho más grande que el de máxima parsimonia se normaliza siendo menor por lo que la agrupación toma mayor fuerza aunque la mutación no se ignora por completo.

Una vez explicado el proceso de evaluación y función objetivo, partimos a la explicación de los criterios de cruza, mutación y el punto de paro.

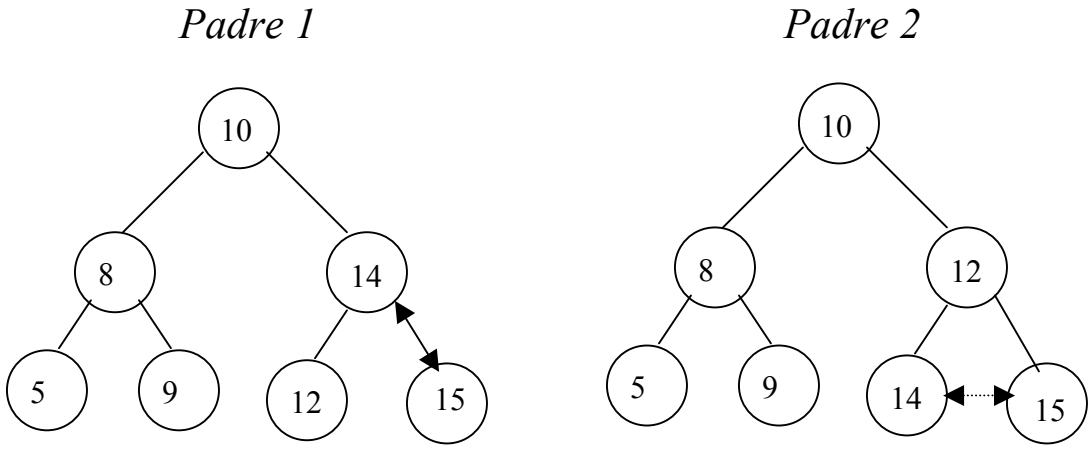
3.2.3 Criterios de Cruza, Mutación y Paro.

Después de tener el valor de la evaluación, procedemos a la cruza de la población. La recombinación el cual tiene una probabilidad del 25% y se realiza mediante el cambio de los elementos de las ramas. Si de manera aleatoria se cumple el 25 % de la probabilidad de cruce, este valor se obtuvo por medio de pruebas ya que con el 25% la población converge de una manera heterogénea y con una probabilidad mayor, los árboles generados se destruyen rápidamente y el error cuadrático medio llega al criterio de paro rápidamente y en el caso de porcentajes menores, tarda mucho en llegar a la convergencia. Se seleccionan los elementos a cruzar por evaluación Montecarlo. Aquí se utiliza una validación para que no existan dos elementos iguales en el mismo árbol.

En forma más detallada, el algoritmo para el cruce es de la siguiente forma:

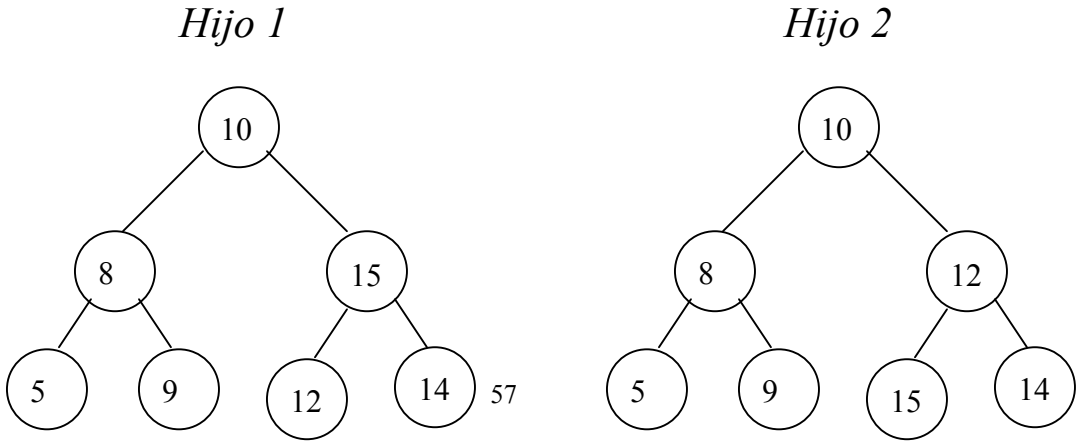
Figura 3.2.7: Algoritmo de cruce de secuencias.

- *Asignar un valor aleatorio entre 0 y 100 a una variable.*
- *Si el valor aleatorio es mayor a 25 entonces.*
    - *Seleccionar de manera aleatoria al primer elemento padre.*
    - *Seleccionar de manera aleatoria el segundo elemento padre validando que no sea igual al primero.*
    - *Seleccionar de manera aleatoria el primer punto de cruce.*
    - *Seleccionar de manera aleatoria el segundo punto de cruce.*
    - *Intercambiar el orden del primer nodo que se eucnetre en el primer padre con el nodo que se encuentre en el segundo padre y validar que no se repita.*
    - *Terminar.*
- *Si no.*
    - *Terminar.*



Si los puntos de cruce son para el primero el 14 y para el segundo el 15.

Se intercambian los valores de los nodos por la posición en la que se encuentran para quedar de la siguiente manera:



En el caso de la mutación, el efecto que produce es un cambio de la posición de los nodos del árbol para generar nuevos árboles (elementos de la población). La probabilidad de que ocurra una mutación es del 1%. Este valor es el valor normalmente utilizado en los algoritmos genéticos y se sabe que produce resultados aceptables, debido a que no se pueden realizar cambios de manera constante, sino que cuando está llegando a un mínimo local, el aplicar una mutación es una forma de poder salir del mínimo local.

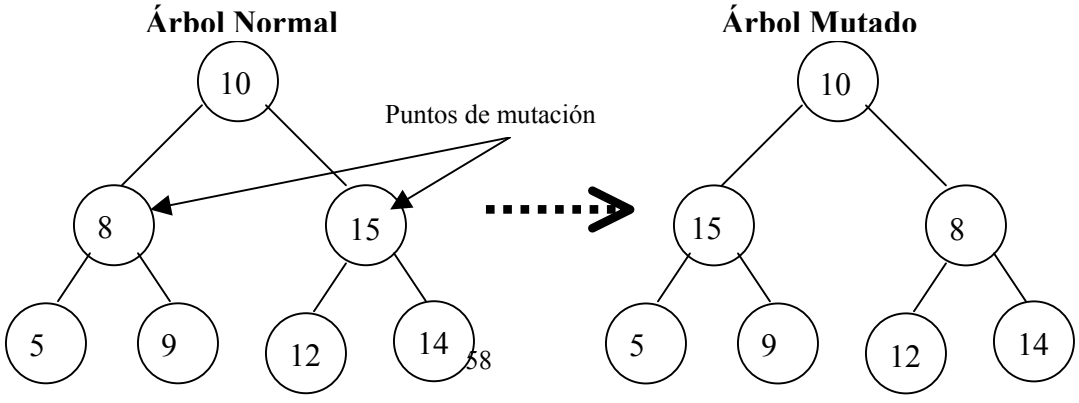
La función de cruza toma un elemento de cada árbol y busca la posición donde se encuentra el valor en el otro árbol para intercambiar el valor en ese lugar. Es decir, se selecciona un nodo del primer árbol padre, se busca ese elemento en el segundo árbol padre para obtener la posición en donde se encuentra y se realiza el intercambio de posiciones entre los elementos del primero con el segundo.

El algoritmo es el siguiente:

Figura 3.2.8 Algoritmo de mutación de secuencias.

- *Asignar un valor aleatorio entre 0 y 100 a una variable.*
- *Si el valor aleatorio es mayor a 99 entonces.*
    - *Seleccionar de manera aleatoria el primer punto de mutación.*
    - *Seleccionar de manera aleatoria el segundo punto de mutación.*
    - *Intercambiar el orden del primer nodo que se eucnetre en el primer punto con el nodo que se encuentre en el segundo punto y validar que no se repita.*
    - *Terminar.*
- *Si no.*
    - *Terminar.*

De una manera gráfica lo podemos ver de la siguiente forma:



Y se repetirá el proceso hasta cumplir alguna de las condiciones de paro.

Esta es la funcionalidad del algoritmo y al final regresará el mejor elemento de la población que será el mejor árbol filogenético de la evaluación heurística.

Posteriormente al algoritmo genético sigue la reconstrucción del árbol en formato Newick.

3.2.4 Reconstrucción del árbol.

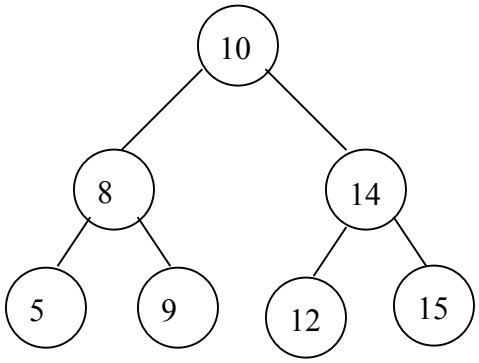
La salida se encuentra en formato Newick en un archivo de formato texto plano que contiene el mejor árbol filogenético híbrido, el mejor del método de Parsimonia Máxima y el mejor de UPGMA, estas salidas podrán ser utilizadas para dibujar el árbol con paquetes como Drawtree de Phyllip. [FELSENSTEIN,2004]

El formato Newick es escrito visitando los nodos y revisando la distancia o diferencia de puntuación que exista entre cada nodo. Si la diferencia es muy grande con respecto a la evaluación promedio del árbol. Entonces se cerrará el paréntesis, de lo contrario, se seguirá en la misma línea. Cada nodo produce tres entradas, es decir:

Figura 3.2.9 Algoritmo del Formato Newick

- *(subarbol_izquierdo,raíz,subarbol_derecho)*
- *visitar raiz de los subarboles.*
- *((subarbol_izquierdo,raíz_subarbol,subarbol_derecho),raiz, (subarbol_izquierdo,raíz_subarbol,subarbol_derecho))*
- *Todo subarbol puede ser un nodo termina, un nulo u otro árbol.*

Para el siguiente árbol el formato newick del árbol sería:



((5,8,9),10,(12,14,15))

4. Experimentos y Resultados.

4.1 Experimentos con HIV

Chin-Yih Ou et al, describen en su artículo “Molecular Epidemiology of HIV Transmission in a Dental Practice” que la transmisión del virus de inmunodeficiencia humana tipo 1 (HIV – 1) a los trabajadores en la salud desde los pacientes está bien documentada, mientras que en el caso contrario no existe mucha información. Ellos comprobarón que los trabajadores en la salud, pueden transmitir el VIH hacia los pacientes, y la relevancia de poder encontrar relaciones filogenéticas en individios de la misma especie, permite establecer rutas de contagio entre las especies para poder realizar estos experimientos es importante el poder tener una relación certera entre las especies evaluadas y la velocidad de procesamiento y tiempo de respuesta, esto fue lo que me motivó para utilizar cadenas de VIH para la evaluación del sistema, además de que existe una separación en la especie y que permite evaluar el resultado. [OU,1992]

Para este experimento, se utilizaron secuencias de ADN previamente alineadas que producen un árbol filogenético sin raíz.

Existen varios tipos de virus del SIDA. Los más frecuentes son el VIH-1 y el VIH-2. Mientras que el VIH-1 se considera responsable de la epidemia que se ha transmitido en el mundo occidental, el VIH-2 parece limitado a la zona oriental del continente africano. La infección por VIH-1 es mucho más agresiva y rápida, comparada con la originada por el VIH-2. Otro tipo de VIH-1 se ha localizado en unas cuantas personas procedentes de Camerún. Si no se indica de forma específica, al referirse a VIH se alude al tipo más prevalente que es el VIH-1.

Para el primer experimento se utilizaron secuencias de VIH alineadas del tipo VIH - 1 y VIH - 2 debido a que la separación de este grupo es conocido y comparable. Las siguientes secuencias de ADN que fueron utilizadas contienen diferentes tipos de secuencias provenientes del virus VIH americano y aficano:

Tabla 4.1.1 Virus de HIV

Número	Nombre
0	Hiv2Ben
1	Hiv2gh1
2	Hiv2d194
3	Hiv2rod
4	Hiv2st
5	Hiv2nihz
6	Hivbru
7	Hivhxb2
8	Hivnl43
9	Hivsf2
10	Hivmn
11	Hivoy
12	Hivrf
13	Hivhan
14	Hiveli
15	Hivndk
16	Hivmal
17	Hiv2z6t

Resultado de Phyllip:

Con el paquete dnapars que es parte del paquete phyllip y que sirve para realizar la parsimonia de árboles filogenéticos con cadenas de ADN tenemos:

((hiv2st:0.04612,(((hivmal:0.05057,((hivndk:0.02427,(hiv2z6t:0.11295,hiveli:0.01961):0.00987):0.02424,(hivoyi:0.02757,((hivhan:0.04023,hivrf:0.03981):0.01099,(((hivmn:0.03241,hivsf2:0.02997):0.01120,(hivn143:0.01560,hivhxb2:0.01132):0.00572):0.05161,hivbru:0.01613):0.01196):0.00721):0.02580):0.03370):0.36273,hiv2nihz:0.05534):0.03351,hiv2rod:0.04007):0.01779):0.04038,(hiv2d194:0.04407,hiv2gh1:0.03678):0.02616,hiv2Ben:0.10081);

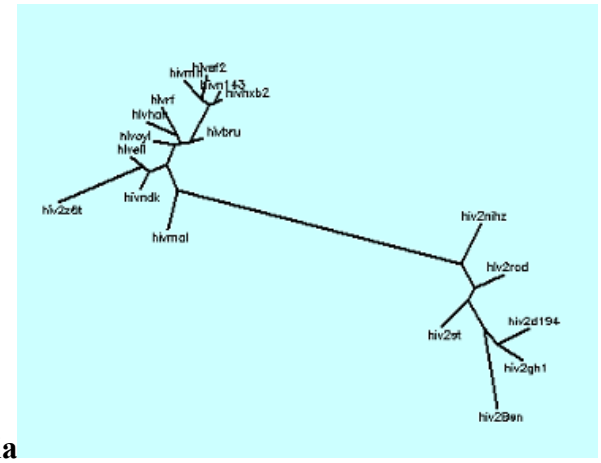
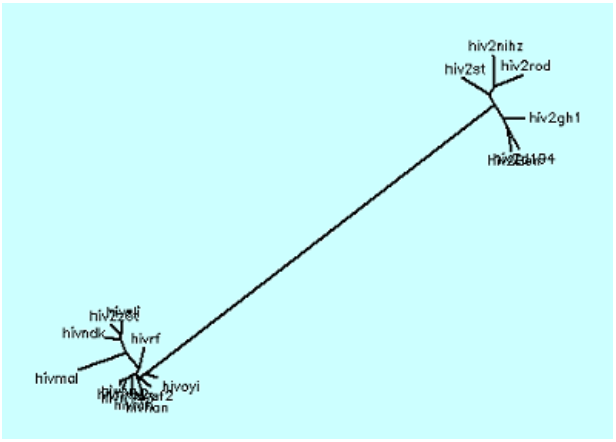


Figura 4.1.1 Árbol Parsimonia

En el caso del DMM con el paquete dnaml:
El formato Newick es:
((((hivoyi:0.01897,(hivsf2:0.02380,(hivhan:0.03773,((hivmn:0.03230,(hivn143:0.01573,(hivhxb2:0.00767,hivbru:0.00470):0.00672):0.01278):0.00627,((hivmal:0.08090,(hivndk:0.02533,(hiv2z6t:0.02383,hiveli:0.02245):0.00774):0.02365):0.03126,hivrf:0.03494):0.00903):0.00272):0.00399):0.01279):0.70205,(hiv2st:0.05013,(hiv2nihz:0.05138,hiv2rod:0.04799):0.01584):0.01891):0.02796,hiv2gh1:0.03282):0.01490,hiv2d194:0.03896,hiv2Ben:0.03778);
Y el árbol es:

Figura 4.1.2 Árbol DMM



En el caso del Híbrido:

El formato Newick es:

(((((((((6 hivbru:112 , 8 hivnl43:383):383 , 15 hivndk:192 , 14 hiveli:401):401 , 13 hivhan:283 , 12 hivrf:256):256 , 11 hivoyi:248 , 10 hivmn:216):216 , 9 hivsf2:475 , 16 hivmal:493):493 , 7 hivhxb2:489):2030,17 hiv2z6t:2030 , 5 hiv2nihz:393 , 3 hiv2rod:386):386 , 4 hiv2st:540 , 2 hiv2d194:301):301 , 1 hiv2gh1:324 , 0 hiv2Ben:0);
Y el árbol es:

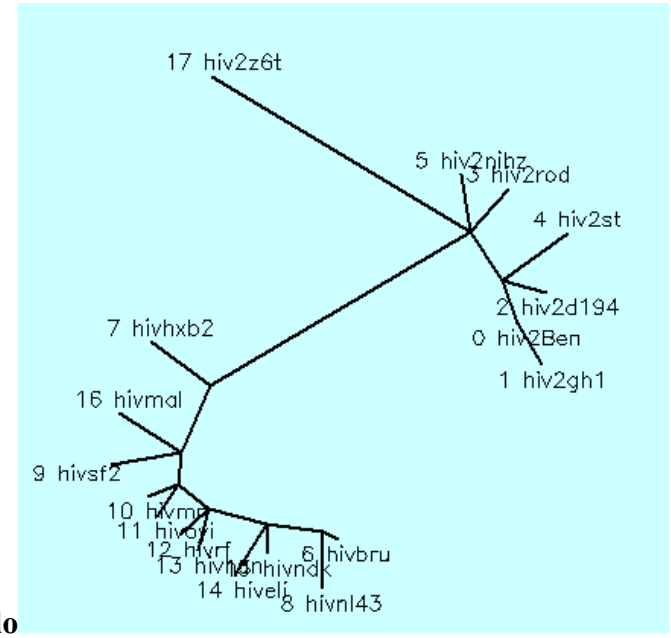


Figura 4.1.3 Árbol Híbrido

En la comparación del árbol híbrido contra el de Phylip en Parsimony se puede ver que la diferencia se da entre la distancia entre los nodos y la estructura de unión de los nodos es similar. Como se observa en los tres árboles, existe una separación grande entre el virus tipo africano HIV-2 con el virus americano HIV-1. Ambos virus se agrupan en extremos del árbol sin raíz.

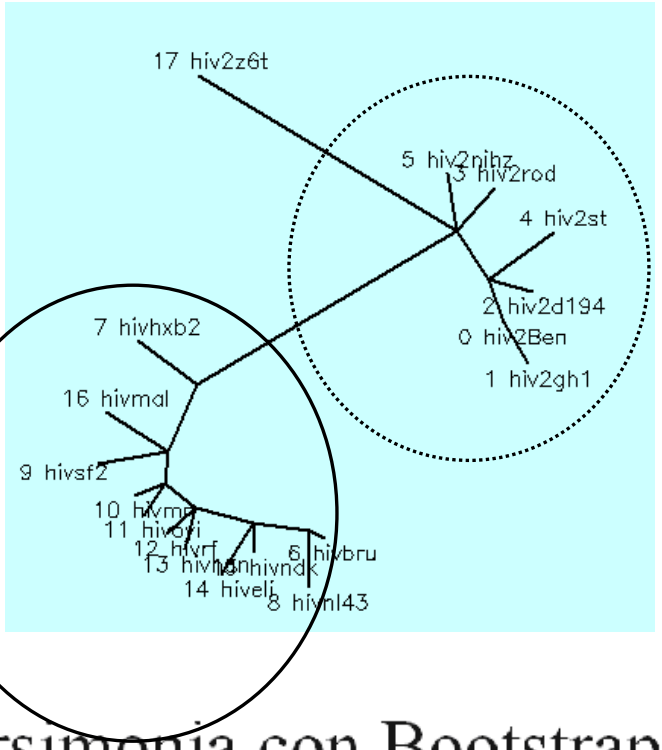
El *bootstrap* sirve para revisar si estadísticamente que árbol tiene mayor probabilidad por lo cual establece el árbol con mayor probabilidad como el más apto.

La única diferencia observable es que en ambos casos que genera el paquete PHYLIP (en DNAPARSE y DNAML), existe un error de clasificación, este error de clasificación es en HIV2Z6T. Que está comprobado que pertenece al virus africano y está clasificado en el grupo del virus americano. Para poder revisar el error de clasificación se optó por revisar en bootstrap que es una validación estadística realizando varias corridas de los elementos de una especie. La clasificación del VIRUS HIV-2 fue divulgada en el año 2002 y se estableció como un subtipo o grupo. Esta decisión fue tomada por el Comité de Nomenclatura del VIH debido a que los virus de este clado son distantes del grupo M (primario) ,N (especie particular distinta a O y M) y O (expecie de afloramiento) del HIV-1, además de que la transmisión hacia los humanos de HIV-2 se suguiere que se dio de una manera independiente a la del HIV-1.

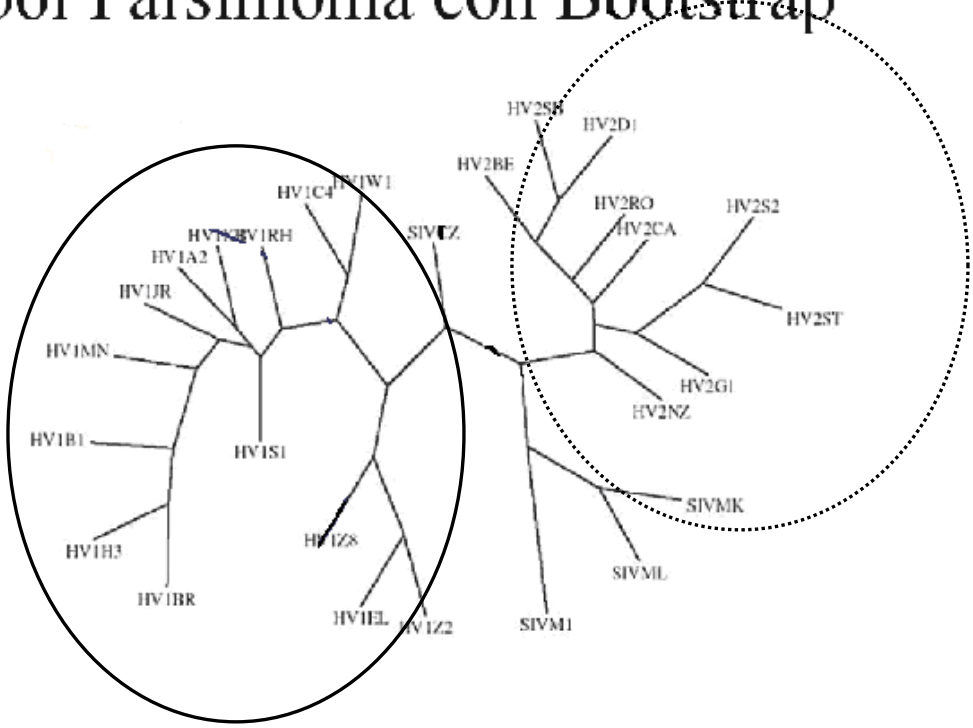
Por esta razón, podemos verificar que el el sistema clasifica de manera adecuada y mejor de PHYLIP. Pero para poder fundamentar más la solución se realizó la segunda validación que se puede ver de la siguiente manera:

Figura 4.1.4 Comparación de Árboles

Árbol Híbrido



Arbol Parsimonia con Bootstrap



Los círculos continuo y punteado muestran la separación parecida entre los árboles híbrido y el validado por *bootstrap* el cual es un método estadístico para comprobación de número de coincidencias de las ramas en formación de árboles con las mismas cadenas.

El primer árbol se obtuvo del sistema creado y el *bootstrap* de un laboratiorio dedicado a estudiar la evolución del HIV y SIV (human y simian virus) este sitio cuenta con validación del gobierno de los Estados Unidos (www.hiv.lanl.gob).

El árbol híbrido utiliza secuencias de DNA mientras que el *bootstrap* utiliza proteínas de *HIV*.

Ambos árboles son similares y el árbol híbrido es una fusión entre DMM y parsimonia en su totalidad generando un árbol desenraizado y es también un árbol bastante lógico separando los *HIV2* del tipo *HIV* por lo cual parece ser un buen método.

Con el método *bootstrap* de validación, podemos observar que las especies de HIV-1 y HIV-2 se separan completamente debido a que está comprobado la cercanía de evolución de estos virus.

4.2 Experimentos con E-Coli.

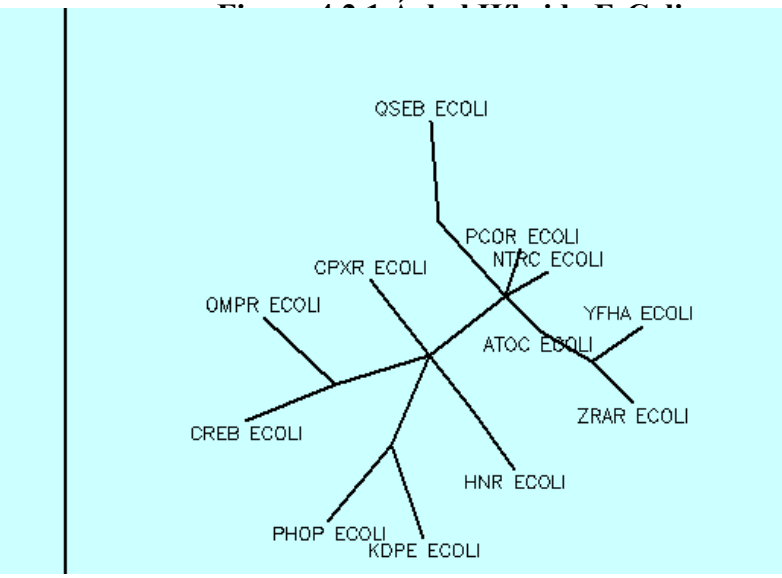
También se realizó otro experimento para validar la correcta funcionalidad del sistema ya que el virus de HIV tiene un tipo de mutación (2-3) que significa que el ADN puede mutar durante el transcurso de su vida, lo que origina que se puedan realizar pruebas equivocadas. En solución a esto se utilizó E. Coli para realizar las pruebas y están avaladas por el experto en biología molecular Dr. Enrique Morett, quién validó la salida del sistema. Las salidas son:

Los dos tipos de e-coli son el un fragmento de la parte de respuesta al medio del E-coli en la parte C y B.

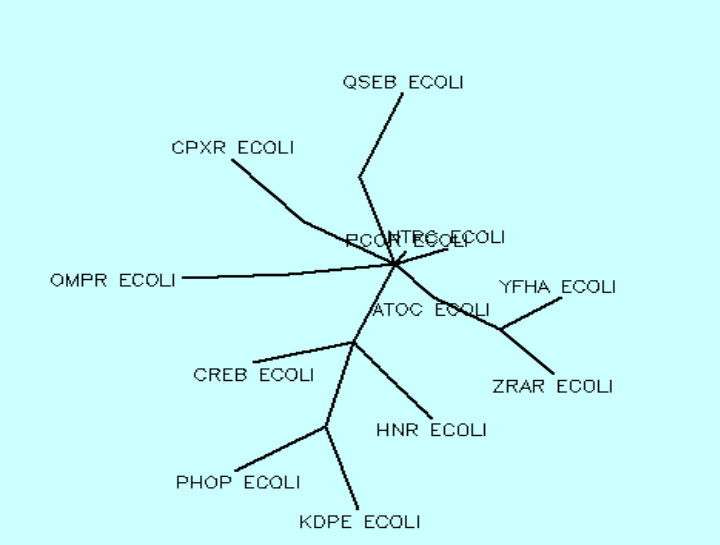
Para el C, las cadenas de entradas de proteina son:

ATOC_ECOLI #--MTAINR-
ILIVDDEDNVRRMLSTAFALQGFETHCANNGRTALHLFADIHPDVVLMDIRMPEMDGIKALKEMRSHETRTPVILMTAYAEVET
AVEALRCGAFDYVIKPFD-
LDELNLIVTGKELIARAIHYNSRRAKGPFIKVNCAALPESLLESELFGHEKGAFTGAQTLRQGLFERANEGTLLLDEIGEMPLVLQ
AKLLRILQEREFERIGGHQTIKVDIRIIAATNRDLQAMVKEG|
ZRAR_ECOLI #--
MTHDNIDILVVDDDISHCTILQALLRGWGYNVALANSGRQALEQVREQVFDLVLCDVRMAEMDGIATLKEIKALNPAIPVLIMT
AYSSVETAVEALKTGALDYLIKPLD-
FDNLQATLTGKELVARAIHASSARSEKPLVTLNCAALNESLLESELFGHEKGAFTGADKRREGRFVEADGGTLFLDEIGDISPM
MQVRLLRAIQEREVQRVGSNQIISVDVRLIAATHRDLAAEVNAG|
YFHA_ECOLI #--
MSHKPAHLLLVDDDPGLLKLLGLRLTSEGYSVVTAESGAEGLRVLNREKVDLVISDLRMDEMDGMQLFAEIQKVQPGMPVIIL
TAHGSIPDAVAATQQGVFSFLTKPVD-
KDALYQAITGKEIFAQAIHNASPRNSKPFIAINCGALPEQLLESELFGHARGAFTGAVSNREGLFQAAEGGTLFLDEIGDMPAPLQ
VKLLRVLQERKVRPLGSNRDIDINVRIISATHRDLPKAMARG|
NTRC_ECOLI #----
MQRGIVWVVDDDSSIRWVLERALAGAGLTCTTFENGAEVLEALASKTPDVLLSDIRMPGMDGLALLKQIKQRHPMLPVIIMTA
HSDLDAAVSAYQQGAFDYLPKPFD-
IDEAVALVTGKELVAHALHRHSPRAKAPFIALNMAAIPKDLIESELFGHEKGAFTGANTIRQGRFEQADGGTLFLDEIGDMPLDV
QTRLLRVLADGQFYRVGGYAPVKVDVRIIAATHQNLEQRVQEG|
PCOR_ECOLI #------
MQRILIVEDEQKTGRYLQQGLVEEGYQADLFNNGRDGLGAASKGQYDLIILDVMLPFLDGWQIISALRESGHEEPVLFLTAKDN
VRDKVKGLELGADDYLIKPFD-FTELVARVAATVCTIADMTVDMVRRTVIRSGKKIHLTGKEYVLLELLLQRTGEVLPR-----
----SLISSLVWNMNFDSDTNVIDVAVRRLRSKIDDDFE-PKLIHTVRGAGYVLEIREE--|
QSEB_ECOLI #-------
MRILLIEDDMLIGDGIKTGLSKMGFSVDWFTQGRQGKEALYSAPYDAVILDLTLPGMDGRDILREWREKGQREPVLILTARDAL
AERVEGLRLGADDYLCKPFA-LIEVAARLASNELRHGNVMLDPGKRIATLAGEPLTLKPKEFALLELLMRNAGRVLSR-------
--KLIEEKLYTWDEEVTSNAVEVHVHHLRRKLGSDF-----IRTVHGIGYTLGEK----|
CPXR_ECOLI #------MNKILLVDDDRELTSLLKELLEMEGFNVIVAHDGEQALDLLD-DSIDLLLLDVMMPKKNGIDTLKALR-
QTHQTPVIMLTARGSELDR>aVLGLELGADDYLPKPFN-
DRELVARIGSPTLEVDALVLNPGRQEASFDGQTLELTGTEFTLLYLLAQHLGQVVSR----------
EHLSQEVLGKRLTPFDRAIDMHISNLRRKLPDRKDGHPWFKTLRGRGYLMVSAS---|
OMPR_ECOLI #---
MQENYKILVVDDDMRLRALLERYLTEQGFQVRSVANAEQMDRLLTRESFHLMVLDLMLPGEDGLSICRRLRSQSNPMPIIMVT
AKGEEVDRIVGLEIGADDYIPKPFN-PRELLARIEEAVIAFGKFKLNLGTREMFREDEPMPLTSGEFAVLKALVSHPREPLSR-------
-------DKLMNLARGREYSAMERSIDVQISRLRRMVEEDPAHPRYIQTVWGLGYVFVPDGSKA|
CREB_ECOLI #----
MQRETVWLVEDEQGIADTLVYMLQQEGFAVEVFERGLPVLDKARKQVPDVMILDVGLPDISGFELCRQLLALHPALPVLFLTA
RSEEVDRLLGLEIGADDYVAKPFS-PREVCARVPSPVIRIGHFELNEPAAQISWFDTPLALTRYEFLLLKTLLKSPGRVWSR-------
-----QQLMDSVWEDAQDTYDRTVDTHIKTLRAKLRAINPDLSPINTHRGMGYSLRGL----|
PHOP_ECOLI #-------
MRVLVVEDNALLRHHLKVQIQDAGHQVDDAEDAKEADYYLNEHIPDIAIVDLGLPDEDGLSLIRRWRSNDVSLPILVLTARES
WQDKVEVLSAGADDYVTKPFH-IEEVMARMASQVISLPPFQVDLSRRELSINDEVIKLTAFEYTIMETLIRNNGKVVSK-------
--DSLMLQLYPDAELRESHTIDVLMGRLRKKIQAQYP-QEVITTVRGQGYLFELR----|
KDPE_ECOLI #------MTNVLIVEDEQAIRRFLRTALEGDGMRVFEAETLQRGLLEAATRKPDLIILDLGLPDGDGIEFIRDLR-
QWSAVPVIVLSARSEESDKIAALDAGADDYLSKPFG-
IGELQARLPDPLVKFSDVTVDLAARVIHRGEEEVHLTPIEFRLLAVLLNNAGKVLTQ----------
RQLLNQVWGPNAVEHSHYLRIYMGHLRQKLEQDPARPRHFITETGIGYRFML-----|
HNR_ECOLI
#MTQPLVGKQILIVEDEQVFRSLLDSWFSSLGATTVLAADGVDALELLGGFTPDLMICDIAMPRMNGLKLLEHIRNRGDQTPVL
VISATENMADIAKALRLGVEDVLLKPVKDLNRLREMVAAKLLQELQPPVQQVISHCRVNYRQLVAADKPGLVLDIAALSENDL
AFY----------CLDVTRAGHNGVLAALLLRALFNGLLQEQLAHQNQRLPELGALLKQVNHLLRQANLP|

El árbol Híbrido de nuestro sistema es:

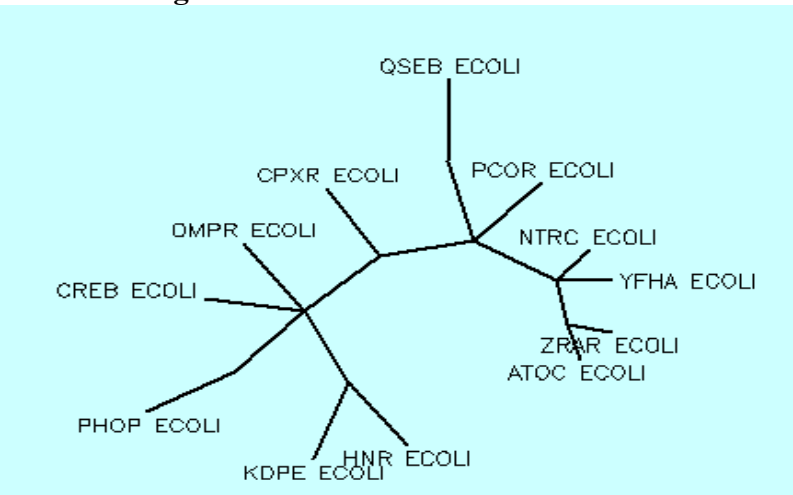


El árbol DMM es:



Y el árbol de Parsomonia es:

Figura 4.2.3 Árbol Parsimonia E-Coli

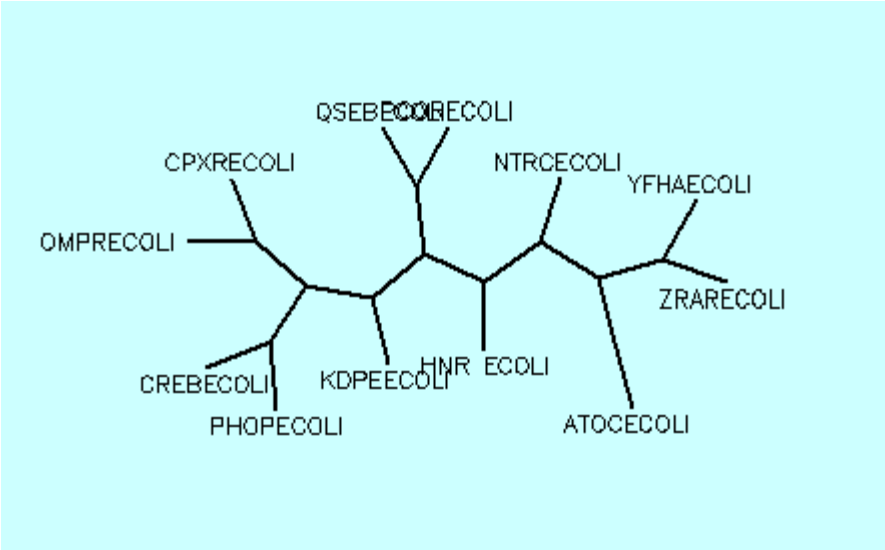


En comparación con Phillip cuyo Newick es:

((((HNR_ECOLI,((KDPEECOLI,((PHOPECOLI,CREBECOLI),(OMPRECOLI,CPXRECOLI))),(QSEBECOLI,PCORECOLI))),NTRCECOLI),(YFHAECOLI,ZRARECOLI)),ATOCECOLI);

Y su árbol es:

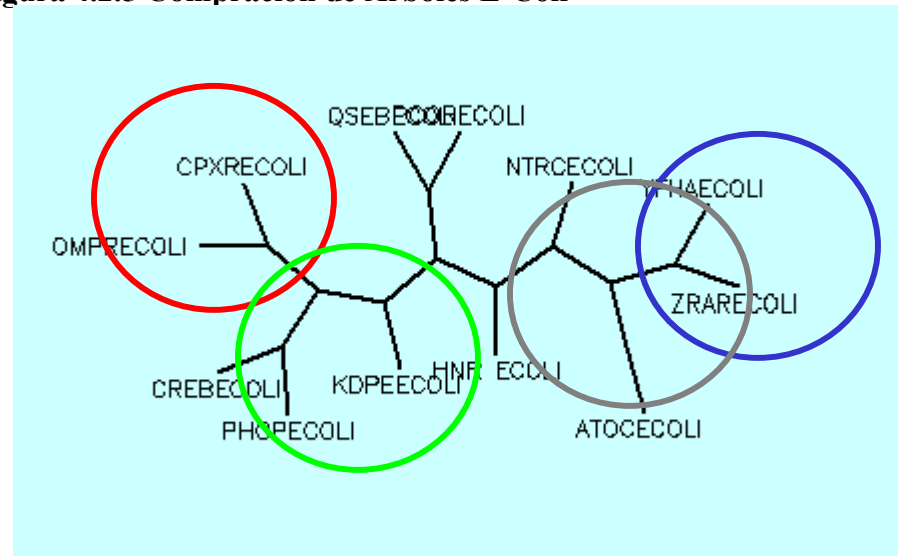
Figura 4.2.4 Árbol Phyllip E-Coli



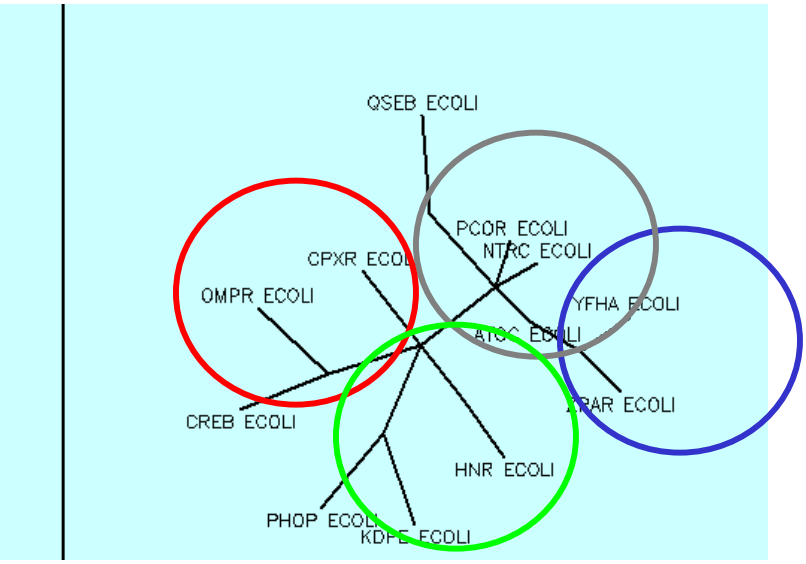
En su similitud tenemos:

El árbol generado por Phylip es:

Figura 4.2.5 Compración de Árboles E-Coli



El árbol generado por nuestro sistema:



Aquí podemos observar que la estructura es similar, ya que agrupa de forma similar, aunque no exactamente igual dado que toma un fusión de los métodos de DMM y Parsimonia. Esta fución refleja la relación de la similitud entre las cadenas con los puntos informativos por lo que varía la distancia entre la unión de las especies, es decir, la separación entre la similitud de las especies.

Las estadísticas que tenemos lo podemos describir de la forma siguiente:

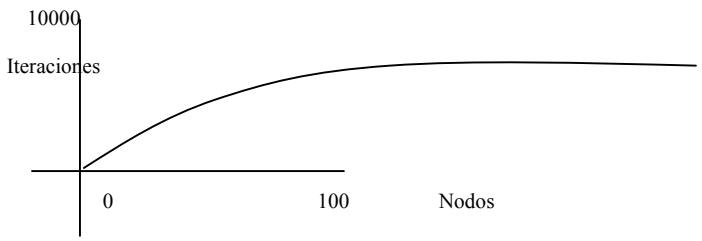
Desempeño: En problemas pequeños, observamos que métodos normales tales como el que utiliza Phyllip son mucho más rápidos, sin embargo, cuando el problema crece, nuestro sistema es mucho más eficiente como se muestra a continuación:

Tabla 4.2.1 Comparación de Rendimiento

Volumen de Procesamiento	Híbrido produciendo los 3 árboles.	Protoparse, DnaParse (Phyllip) produciendo un árbol
HIV 18 cadenas 10650 aminoácidos	33.67 SEG.	1 MIN. 10.50 SEG.
Ecoli-C 15 cadenas 362 proteínas	2.54 SEG.	1.43 SEG.
Ecoli-B 12 cadenas 486 proteínas	2.98 SEG.	2.01 SEG.

Convergencia del Algoritmo Genético: La convergencia es rápida debido a la estrategia de conservación del mejor evaluado (elitismo) explicado en el capítulo anterior. El mejor pasa automáticamente a la siguiente generación, garantizando que siempre tienda al mejor.

Es por esto que la convergencia del algoritmo genético tiene un comportamiento logarítmico:

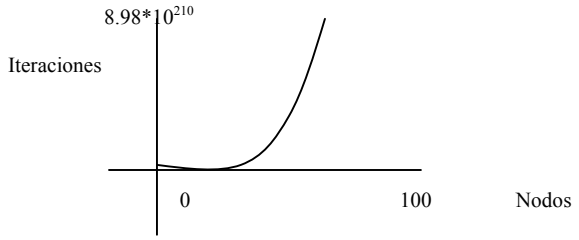


En nuestro caso, el mejor de los casos es de 10 iteraciones, el primedio es de 163 iteraciones y el peor caso es del criterio de paro de 10000 iteraciones.

Mientras que en el caso clásico, la complejidad sigue la ecuación:

$$NN = \frac{(2n-5)!}{[2n-3(n-3)!]}$$

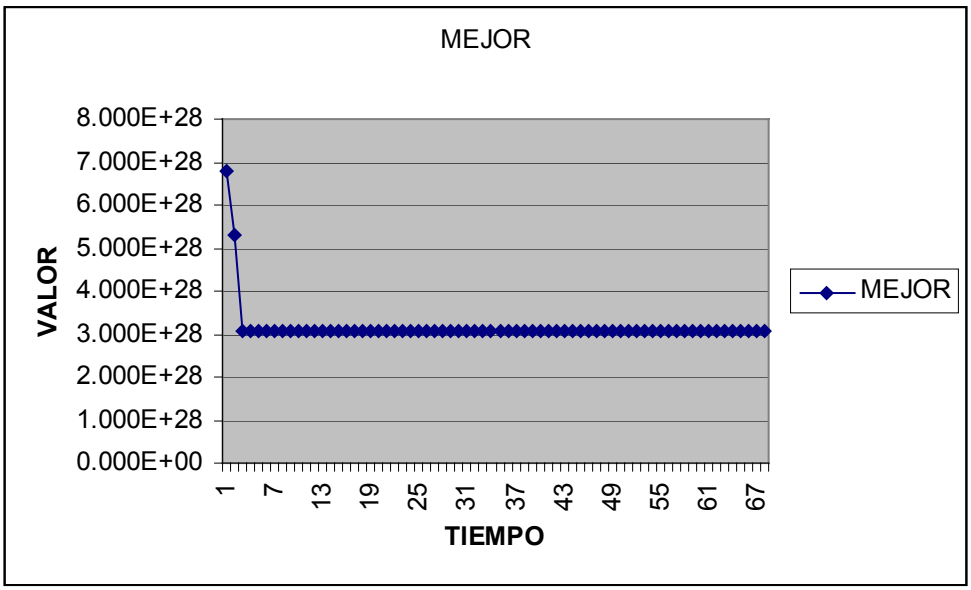
Por lo que su crecimiento es exponencial:



Como podemos ver en el comportamiento, las ventajas que se tienen con este sistema es que se pueden producir los tres tipos de árboles de manera rápida y eficiente.

En cuanto al rendimiento del Algoritmo tenemos la siguiente gráfica que muestra el mejor, el peor y el promedio de los casos según el tiempo:

Debido a que el valor del mejor en comparación con el promedio y el peor es muy estable en la gráfica anterior no se puede ver el comportamiento por lo que se muestra en la siguiente gráfica y tabla:

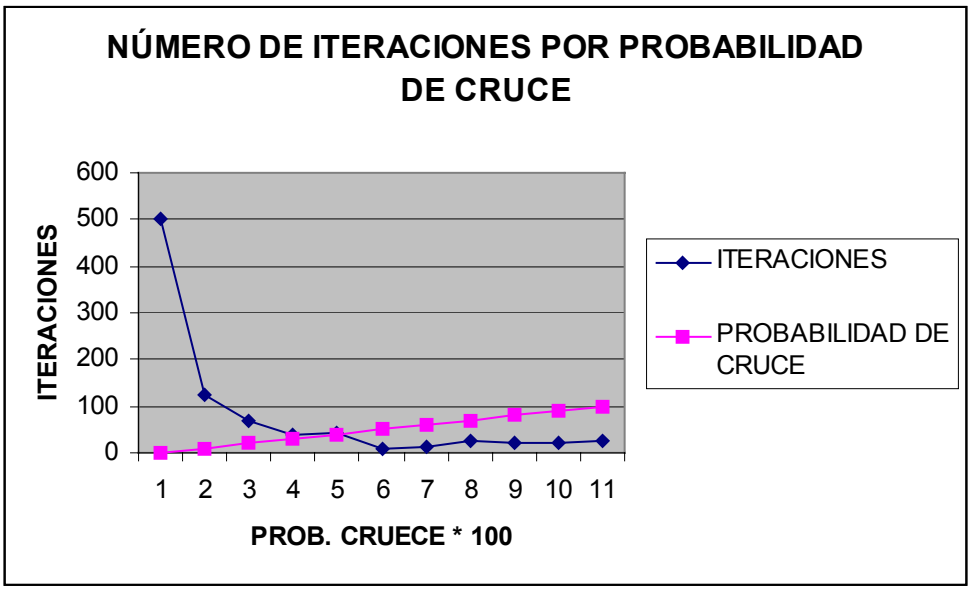


6.770974401783110E+28
5.317573558296910E+28
3.074052798186080E+28
3.074052798193960E+28
3.074052798197670E+28
3.074052798205100E+28
3.074052798205100E+28
3.074052798219970E+28
3.074052798205100E+28
3.074052798205100E+28
3.074052798219970E+28
3.074052798223680E+28
3.074052798227400E+28
3.074052798231110E+28
3.074052798234830E+28
3.074052798238540E+28
3.074052798238540E+28
3.074052798242430E+28
3.074052798246140E+28
3.074052798246140E+28
3.074052798253570E+28
3.074052798257290E+28
3.074052798261010E+28
3.074052798264720E+28
3.074052798268440E+28

3.074052798275810E+28
3.074052798275810E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798279530E+28
3.074052798338810E+28
3.074052798338810E+28
3.074052798338810E+28
3.074052798338810E+28
3.074052798338810E+28
3.074052798383950E+28
3.074052798383950E+28
3.074052798383950E+28
3.074052798383950E+28
3.074052798383950E+28
3.074052798383950E+28
3.074052798383950E+28
3.074052798383950E+28
3.074052798383950E+28
3.074052798383950E+28
3.074052798383950E+28

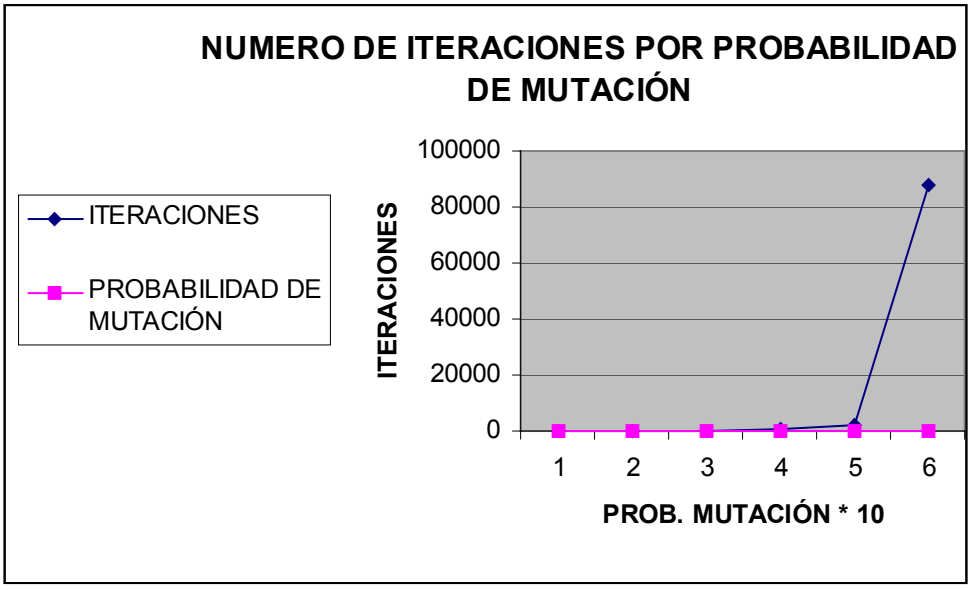
Para evaluaciones de selección de la mejor opción tenemos:

ITERACIONES	PROBABILIDAD DE CRUCE
500	1
123	10
69	20
39	30
44	40
8	50
12	60
24	70
21	80
22	90
25	100



Esta estadística y la siguiente tabla muestran el número de mutaciones promedio por iteraciones o generaciones de árboles de la metodología utilizada.

ITERACIONES	PROBABILIDAD DE MUTACIÓN
8	1
88	10
104	20
500	30
2500	40
87950	50



Y para la selección del valor para cruce y para mutación obtenemos de las gráficas los valores para la probabilidad de cruce el valor es del 60 % y para la mutación del 1 %.

La heurística produce un resultado que es una fusión que explota el potencial de dos métodos que son utilizados actualmente, y que se conoce que produce buenos resultados. La fusión de ambos métodos para generar el árbol es una forma de explotar las fortalezas que tienen ambos métodos, DMM toma en cuenta la similitud entre las cadenas en toda su magnitud, por lo que la mutación toma un papel total en la evaluación, por lo que una mutación o algún cambio pequeño también entrará en consideración, por el otro lado, *Parsimony* solo toma en cuenta los sitios informativos, los cuales estaría ignorando toda la cadena restante y es por eso que se pueden generar varios árboles que sean los mejores.

El resultado emite resultados cuya agrupación es mucho mejor que con solo un elemento como mostramos con los experimentos.

5. Conclusiones y Trabajos Futuros.

5.1 Conclusiones

Este documento describe un método para la construcción de árboles filogenéticos que considera dos criterios para la generación de árboles que son el *Distance Matrix Method (DMM)* y el *Parsimony Method* y por medio de la combinación de los métodos proporcionar un enfoque más completo para la clasificación.

El modelo propuesto es la fusión de dos métodos para generar árboles filogenéticos y es una forma de explotar las fortalezas que tienen ambos métodos ya que DMM toma en cuenta la similitud entre las cadenas, por lo que toda la secuencia entrará en consideración, por el otro lado, Parsimonia solo toma en cuenta los sitios informativos, y es así que ignora toda la secuencia restante. Teniendo así un método de evaluación heurística robusta que genera un solo árbol apto y mejor evaluado.

En nuestro sistema híbrido solo un árbol es tomado como el mejor, además de que se toma en cuenta la evaluación de las partes informativas como la similitud, y poder así generar un árbol que explote esta ventaja.

Esta forma de hibridación de métodos no es sólo la mejora desde el punto de vista de utilizar las ventajas de ambos métodos, sino que la velocidad es rápida para generar los árboles para poder comparar y evaluar la mejor salida, ya que se utiliza heurística en vez de un cálculo exhaustivo. Todo esto se puede observar en las tablas del inciso anterior.

De esta manera se comprueba que la aplicación de algoritmos evolutivos tales como los algoritmos genéticos con una función objetivo híbrida (que contiene más de un criterio de

evaluación) produce clasificaciones eficientes, además de que la velocidad de respuesta es mucho más rápida que los métodos clásicos.

Para esta investigación, se utilizó algoritmos genéticos cuya función objetivo constó de dos criterios de evaluación, máxima parsimonia y método de matriz de distancias, en el caso de máxima parsimonia su contribución en la función objetivo fue del 100 % mientras que para la matríz de distancia fue de 2.4% debido a la teoría neutralista que establece que solo el 2.4% de las mutaciones tiene una fijación en la evolución. Con esto se logra un equilibrio para tomar en cuenta la mutación solo en un porcentaje y así mismo encontrar los puntos en los que las especies se separan. Para el criterio de cruza, se encontró que la probabilidad del 60% es el que produce mejores resultados junto con una probabilidad de mutación del 1%.

La importancia en los resultados de la investigación es que se demostró que con un método híbrido se puede explotar importantes beneficios que suplan las deficiencias entre los métodos. Como en el caso de HIV, se puede ver una clara separación que comprueba que un solo método es a veces no muy eficiente mientras que la fusión de dos métodos realizan una mejor labor.

Como otro aspecto importante se puede mencionar que la hibridación para la evaluación no solo tiene que ser de dos métodos, sino que se pueden utilizar una amplia gama de combinaciones que logren explotar ventajas de todos los elementos y lograr una busqueda más eficiente, ya que el parámetro que establece la evaluación para la búsqueda es la función objetivo. Esta hibridación no sólo se podrá utilizar en la construcción de árboles filogenéticos sino en muchas otras búsquedas heurísticas.

Las limitantes que se tienen es que algunas veces una hibridación pueda limitar a un método como en el caso de matriz de distancia con parsimonia, esto se refleja en que sólo se conseguirá un árbol el cual es considerado el más apto, aunque el método de parsimonia pueda generar más de un árbol que cumpla con las condiciones, matriz de distancia lo limitará a solo una posibilidad.

5.2. Trabajos Futuros

Este método fue probado con base en comparaciones con otros métodos que se han comprobado que producen un resultado aceptable, por lo que se pueden seguir realizando evaluaciones para validar el método. Esto utilizando diferentes ADN y proteínas de organismos diferentes.

Nuestra investigación sólo abarca dos métodos los cuales se complementan, sin embargo, se pueden utilizar varios métodos de forma híbrida para poder así generar otro método. Con lo cual se puede generar árboles que tengan las características de los potenciales de los otros modelos en cuanto al punto de vista cladista y fenetista, el cual tomará la evolución cercana a la real.

Para estas pruebas se contó con el apoyo de expertos en biología molecular, todavía se puede profundizar en la evaluación de este método.

Existen también otras heurísticas las cuales podrán ser aplicadas con una evaluación híbrida o multiple, tales como recocido simulado, programación genética o cualquier otro algoritmo evolutivo.

Es importante comentar que este método podrá ser utilizado para realizar investigaciones de epidemiología, realizando adecuaciones y verificando los criterios de cruza, mutación, paro y pesos en la función objetivo.

También se puede realizar una búsqueda más eficiente si se pudiera dirigir la búsqueda aplicando métodos como las cadenas de Markov para poner probabilidades e infactibilidad en las filogenias que se conozcan de antemano que son improbables o poco probables.

Realizar un análisis matemático formal sobre la generación, aunque esté basado en heuristicas, se pueden realizar validaciones sobre las convergencias y la utilización del método combinado.

Realizar estudios con organismos de diferentes especies para poder obtener una relación no solo intraespecie sino interespecie.

Por último se hace incapié en que las heurísticas con evaluación híbrida no sólo se pueden aplicar a la generación de filogenias, sino que cualquier otro tipo de búsqueda que se pueda realizar con varios métodos y que cada uno tenga una ventaja podrá ser utilizada y crear una clasificación o búsqueda más eficiente.

Glosario:

Ácido Nucleico: Molécula polimérica que contiene información genética como base de secuencias.

Adenina, Timina: Guanina, Ciotisina — Las cuatro bases que conforman el ADN.

Amino Ácido: Monomero del cual las proteínas están formadas.

ADN: Llamado también Ácido DisoxiRibo Nucleico, es el ácido nucléico del polímero, constitutivo de los genes.

Alelo: Versión particular de un gene.

Árbol o cladograma: Una manera de presentar la filogenia de los organismos; se ha preparado un árbol o cladograma de los grupos de organismos estudiados en este curso según el website del Árbol de la Vida

ARN: También llamado ácido, es otro ácido nucleico que difiere del ADN en que tiene un azúcar ribosa en vez de disoxiribosa.

ARN Mensajero: Clase de ARN molecular que utiliza el ADN como mensajero para llevar ordenes al resto de las células, es abreviado como mARN.

ARN Polimeraza: Enzima que sintetiza el ARN utilizando el molde ADN.

ARN transferencia: Moléculas ARN que acarrean el amino ácido a los ribosomas.

Bioinformática: Ciencia que trata los problemas de la biología molecular utilizando técnicas computacionales.

Cladística: La clasificación filogenética de los organismos basada en los caracteres sinapomórficos

Clado o Clade (*f*): Una rama en un árbol filogenético; los miembros de un clado son caracterizados por caracteres sinapomórficos, excepto cuando han ocurrido reversiones

Codón: Grupo formado por 3 bases de ARN o DNA el cual codifica un amino ácido.

Cromosoma: Estructura de un gen de una célula hecha de una molécula de ADN.

Convergencia: La adquisición de caracteres similares en líneas evolutivas distintas, típicamente bajo condiciones ambientales o presiones selectivas similares

Diploide, Triploide y Tetraploide	Que tiene dos, tres o cuatro copias de cada gen respectivamente.
Enzima:	Proteína que acarrea una reacción química.
Epítasis:	Cuando la mutación en un gen cubre el efecto de alteración en otro gen.
Estándar: Complementario	Dos es estándares son complementarios si A en una es siempre pareja de T en la otra y G es siempre pareja de C o viceversa.
Exón:	Segmento del gen que codifica a la proteína.
Fenotipo:	Efecto visible del genotipo.
Filogenia:	La historia evolucionaria de los taxones o grupos de organismos biológicos
Filogenia molecular:	La clasificación filogenética de los organismos según sus características moleculares, típicamente basada en las secuencias de los aminoácidos en las proteínas o en las secuencias de los nucleotidos en el ácido desoxirribonucleico que se encuentra en los plastidios, las mitocondrias y el genoma nuclear
Gen:	Unidad de información genética.
Genotipo:	La descripción total de un organismo.
Haploide:	Qué tiene una copia de cada gen.
Homoplasia:	La presencia de caracteres similares, adquiridos independientemente, en líneas evolutivas distintas
Intrón:	Segmento del gen que no codifica a la proteína.
Mutación:	Alteración de la información Genética.
Mutación : Silenciosa	Mutación cuyo efecto no afecta en la sobre vivencia y crecimiento de la célula y no tiene efectos observables.
Proteína:	Polímero hecho de aminoácidos, ellos realizan casi todas las funciones de la célula y definen su estructura.

Puente de : Hidrógeno	Conexión resultante de la atracción de loa átomos positivos del hidrógeno con los átomos de carga negativa.
Replicación:	Duplicación del ADN posterior a la división celular.
Reversión:	La pérdida de un carácter derivado y el regreso al carácter ancestral
Sinapomorfia:	Un carácter derivado compartido entre dos o más de dos taxones
Trascripción:	Proceso por el cual la información del ADN es convertida en su equivalente de ARN.
Traducción:	Creación de proteínas utilizando la información proveniente del ARN mensajero.
Uracilo:	Base que reemplaza la timina en la molécula de ARN y que puede hacer pareja con la adenina.

BIBLIOGRAFÍA:

Clark David & Russell Lonnie, *Molecular Biology made simple and fun*, EUA, Cache River Press, 2000

Pevzner Pavel A., *Computational Molecular Biology An Algorithmic Approach*, Cambrige Massachusetts EUA, The MIT Press, 2000.

Dan Graur y Wen-Hsiung Li, *Fundamentals of Molecular Evolution*, Sunderland Massachusetts EUA, Sinauer Associates, INC., 2000.

http://www.ncbi.nlm.nih.gov/About/primer/phylo.html

http://www.blues.uab.es/~icgm4/tema5/tsld007.htm

Rafael Martí, *Algoritmos Genéticos*, www.cag.com.mx/Algoritmos Genéticos.htm,2000

Monserrrat Aguadé, Evolución molecular: el reloj de la vida, Barcelona España, www.uv.es/metode/anuario2000/167_2000.html

Bates Cogdon Clare, Gaphyl: An Evolutionary Algorithms Approach fro the Study of Natural Evolution, EUA,Colby Collage,2002

T. H. Reijmers, et. Al, Using genetic algorithms for the construction of phylogenetic trees: application to G-protein coupled receptor sequences , Laboratory of Analytical Chemistry, University of Nijmegen, 6525 ED Nijmegen, The Netherlands y Center for Molecular Design, Janssen Research Foundation, B-2350 Vosselaar, Belgium 1998.

MEDLINE, http://www.ncbi.nlm.nih.gov/,USA,2004

Motoo Kimura , "The neutral theory of molecular evolution", Cambridge University Press.1983, reimpresión 1986

Roderic Pages, *Introduction to Tree Building*, University of Glasgow, 2000

Ou Chin-Yih, et. Al, Molecular epidemiology of HIV transmission in a dental practice, Centers for Disease Control, Atlanta USA, 1992

Joe Felsenstein, PHYLIP, Department of Genome Science, University of Washington USA, 2004