



ANÁLISIS DE ALGORITMOS PARA LA GENERACIÓN DE FILOGENIAS

T E S I S

QUE PARA OBTENER EL GRADO DE:
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

P R E S E N T A

Benito Samuel López Razo

DIRECTOR DE TESIS

Dr en Ed. JOEL AYALA DE LA VEGA

Tutores Adjuntos

Dr.en C. Oziel Lugo Espinosa

Dr. en C. Alfonso Zarco

Diciembre, 2016



DICTÁMEN DE AUTORIZACIÓN Y OBTENCIÓN DE GRADO DE MAESTRÍA

Texcoco, Méx. , a 09 de Noviembre del 2014

Título del proyecto:

ANALISIS DE ALGORITMOS PARA LA GENERACIÓN DE FILOGENIAS

Tesista:

Benito Samuel López Razo

Dictamen:

No. de revisión: 3

- Rechazado
- Sujeto a modificaciones
- Aceptado, condicionado
- Aceptado



Observaciones generales:

Acceptedo para la impresión

Acceptedo para la defensa de grado

<p>Tutor Académico</p>  <p>Dr. en Ed. Joel Ayala de la Vega</p>	<p>Tutor Adjunto</p>  <p>Dr. eg.C. Oziel Lugo Espinosa</p>	<p>Tutor Adjunto</p>  <p>Dr. en C. Alfonso Zarco Hidalgo</p>
--	---	---



RESUMEN

Usualmente la generación de árboles de reconstrucción filogenética se realiza considerando información examinada dentro de un conjunto de especies previamente definidas. La información de dichas especies puede ser representada en base a caracteres homólogos o secuencias de ADN.

Para el análisis de datos mediante caracteres homólogos existen herramientas tecnológicas (software) que facilitan este trabajo, sin embargo, muchas de estas herramientas están diseñadas para trabajar bajo ciertas características, características que si no se cumplen es imposible el acceso a estos recursos.

En este trabajo se desarrolla un sistema de información que sea de código libre e independiente a plataforma, el cual se basa en el uso de caracteres homólogos para analizar el conjunto deseado de especies y utilizar el algoritmo de Hennig como base para la reconstrucción de filogenias además del uso del algoritmo Simple LinkAge como una alternativa para dicha reconstrucción. Una vez concretado el sistema, es viable su traslado a una plataforma web que pueda ser consultada en línea.

Agradecimientos

A Dios por todas las bendiciones recibidas. Porque la vida como la historia se dividen antes y después del Creador.

A mis padres Benito López Pérez y Martha Razo Arellanes por todo su cariño, cuidado y por jamás dejarme caer. Por enseñarme a ser fuerte cuando más débil soy

Al Dr. Joel Ayala de la Vega por ser el capitán de este barco y el guía de este viaje. Por toda su experiencia, apoyo incondicional y por extender la mano a los alumnos para ser mejores día a día, mi respeto y admiración eternos.

Al Dr. Jesús Romero Nápoles por participar en este proyecto, por su tiempo y los consejos recibidos.

A la Universidad Autónoma del Estado de México, por hacerme sentir orgulloso, importante y por formar profesionales ¡Patria, ciencia y Trabajo, potros gloria!

Al Consejo Nacional de Ciencia y Tecnología por el apoyo económico para realizar estudios de posgrado

Al Consejo Mexiquense de Ciencia y Tecnología por dar apoyo a los alumnos mexiquenses

Al Colegio de Posgraduados por abrir las puertas para la investigación

A las Reales Fuerzas especiales MACSCO 2014 por ser compañeros, amigos y hermanos. Por la ayuda brindada, por las risas, enojos, exámenes que aprobamos juntos.

Dedicatoria

A mi padre, al hombre que más admiro y respeto en el mundo. Por jamás dejarse vencer (Me doblo, pero no me quiebro)

A mi madre, por ser lo más bello que Dios puso en mi vida y a pesar de los días grises siempre tiene un motivo para sonreír

A mis hermanas, por ser mis confidentes y mis más grandes ejemplos

Contenido

Índice de Figuras	VI
Índice de Tablas.....	VII
Prólogo	1
1. Introducción.....	1
1.1 Planteamiento del problema	1
1.2 Justificación.....	1
1.3 Objetivo General.....	2
1.4 Objetivos específicos	3
1.5 Supuesto	3
2. Trabajos Previos y Antecedentes	3
2.1 Definición de Entomología	3
2.2 Orígenes de la Sistemática	4
2.3 Taxonomía y Sistemática.....	6
2.4 La Filogenia como base de la Sistemática.....	6
2.5 Escuelas de la Sistemática	7
2.5.1 Sistemática Evolutiva o tradicional.....	7
2.5.2 Numérica o Fenética.....	8
2.5.3 Filogenética o cladística.....	8
3. Principios Básicos.....	11
3.1 Elementos Dentro de la Cladística	11
3.2 Las funciones de la Sistemática.....	12
3.3 Caracteres	13
3.4 Matrices de Datos.....	14
3.5 Cladogramas.....	16
3.6 Estadísticos descriptivos.....	18
3.6.1 Longitud del Cladograma	18
3.6.2 Índice de consistencia	18
3.6.3 Índice de retención.....	19
3.6.4 Índice de consistencia rescalado.....	19
3.6.5 Grupos monofiléticos	19
3.6.6 Grupos parafiléticos	19
4. Métodos de Inferencia Filogenética	20
4.1 Teoría de Grafos.....	21
4.2 Grado de un grafo	22
4.3 Construcción de Árboles Filogenéticos.....	23

4.4	Raíz de un árbol filogenético.....	24
4.5	Complejidad Matemática	25
4.6	Métodos Basados en Distancias.....	26
4.6.1.	WPGMA (Weighted Pair Group Method Using Arithmetic Average)	26
4.6.2	UPGMA	26
5.	Hennig.....	27
5.1	Diagrama de Flujo.....	29
5.2	Seudocódigo	29
5.3	Complejidad Matemática de Hennig	30
5.4	Metodología de programación Hennig.....	30
6.	Metodología Propuesta Simple LinkAge.....	31
6.1	Algoritmos de partición	32
6.2	Algoritmos jerárquicos	32
6.3	Etapas del Análisis de Conglomerados	33
6.4	Diagrama de Flujo.....	34
6.5	Complejidad Matemática Simple Linkage	35
6.6	Seudocódigo	35
7	Resultados y Funcionamiento del Sistema	37
7.1	Creación de un árbol filogenético en base al algoritmo de conglomerados Simple Linkage	37
7.2	Funcionamiento del sistema	42
	Discusión.....	45
	Conclusiones	46
	Bibliografía	47
	Anexos.....	48
	Artículo en Revista Indexada.....	49
	Capítulo de libro.....	67

Índice de Figuras

Figura 1	Línea de Tiempo de científicos que han participado en el progreso de la evolución	5
Figura 2	(a) Autapomorfía; (b) sinapomorfías; (c) simplisiomorfía	14
Figura 3	(a) Paralelismo; (b) convergencia; (c) reversión.....	14
Figura 4	Partes de un Cladograma	16
Figura 5	Cladogramas dicotómicos	17
Figura 6	Cladogramas Politómicos	17
Figura 7	Cladogramas con raíz y sin raíz.....	18

Figura 8 Cladograma con grupo Parafilético	20
Figura 9 Secuencia de ADN.....	20
Figura 10 Grafo dirigido.....	21
Figura 11 Grado de Incidencia Positivo	22
Figura 12 Grado de Incidencia Negativo	22
Figura 13 Grado de un nodo.....	23
Figura 14 Árbol	23
Figura 15 Árbol con raíz y sin raíz.....	24
Figura 16 Árboles generados con 3 especies	25
Figura 17 Número de árboles con y sin raíz	25
Figura 18 Árbol generado con el carácter 1 (Lipscomb, 1998)	28
Figura 19 Árbol generado con el carácter 2 (Lipscomb, 1998).....	28
Figura 20 Árbol generado con el carácter 3 (Lipscomb, 1998)]	28
Figura 21 Árbol generado con el carácter 4 (Lipscomb, 1998)	28
Figura 22 Árbol generado con el carácter 5 (Lipscomb, 1998)]	29
Figura 23 Diagrama de Flujo Hennig	29
Figura 24 Matriz de Datos ordenada por filas.....	30
Figura 25 Matriz de datos ordenada por filas y columnas	31
Figura 26 Cladograma generado con la aplicación desarrollada	31
Figura 27 Clasificación de algoritmos jerárquicos.....	32
Figura 28 Distancia del vecino más cercano	33
Figura 29 Diagrama de Flujo Conglomerados	34
Figura 30 Conglomerado (Autoría Propia)	36
Figura 31 Conglomerado entre A-B-C (Autoría Propia).....	37
Figura 32 Conglomerado (Autoría Propia)	38
Figura 33 Conglomerado (Autoría Propia)	38
Figura 34 Conglomerado (Autoría Propia)	39
Figura 35 Conglomerado (autoría propia).....	40
Figura 36 Conglomerado (Autoría Propia)	40
Figura 37 Conglomerado (autoría propia).....	41
Figura 38 Conglomerado (Autoría Propia)	41
Figura 39 Árbol filogenético a) creado con el algoritmo de Hennig (Nápoles, 2015) b) creado con el algoritmo de conglomerados	42
Figura 40 Pantalla inicial de la aplicación.....	42
Figura 41 Seleccionar número de Especies y Características.....	43
Figura 42 Interfaz para la captura de información.....	43
Figura 43 Tabla de Datos a ser analizados	44
Figura 44 Datos analizados.....	44
Figura 45 Resultado final.....	44

Índice de Tablas

Tabla 1 Comparación de métodos en la Sistemática.....	9
Tabla 2 Matriz de datos ejemplo.....	15
Tabla 3 Ejemplo de conjunto de taxas	15
Tabla 4 Ejemplo una Matriz de Datos Completa.....	15
Tabla 5 Métodos de reconstrucción filogenética, ventajas y desventajas	27
Tabla 6 Tabla de datos con 5 especies	27

Tabla 7 Matriz de Datos (Autoría Propia).....	35
Tabla 8 Matriz para calcular las diferencias entre especies (Autoría Propia)	36
Tabla 9 Matriz de Distancias (Autoría Propia).....	36
Tabla 10 Matriz de Datos (Autoría Propia).....	36
Tabla 11 Matriz de Distancias (Autoría Propia).....	37
Tabla 12 matriz de Datos (Nápoles, 2015)	37
Tabla 13 Matriz de Distancias (Autoría Propia).....	38
Tabla 14 Matriz de Distancias (Autoría Propia).....	38
Tabla 15 Matriz de Distancias (Autoría Propia).....	39
Tabla 16 Matriz de Distancias (Autoría Propia).....	39
Tabla 17 Matriz de Distancias (Autoría Propia).....	40
Tabla 18 Matriz de Distancias (Autoría Propia).....	40
Tabla 19 Matriz de Distancias (Autoría Propia).....	41
Tabla 20 Matriz de Distancias (Autoría Propia).....	41

Prólogo

El proceso de reconstrucción de filogenias dentro del campo entomológico no es un tema nuevo, de hecho, existen muchas herramientas que ayudan a realizar esta labor. En México, debido a la gran abundancia de fauna se necesitan herramientas que faciliten el cumplimiento de este fin.

En el departamento de Entomología y Acarología del Colegio de Posgraduados el Dr. Jesús Romero Nápoles ha trabajado con numerosas herramientas de software que han facilitado la reconstrucción de árboles filogenéticos. En 1989 J. S. Farris produjo Hennig 86, un programa de parsimonia que incluía búsqueda branch and bound para la búsqueda del árbol más parsimonioso, se ejecuta en microcomputadoras compatibles con PC con al menos 512K de memoria RAM y no necesita un coprocesador matemático o los gráficos del monitor (Washington, s.f.).

1. Introducción

1.1 Planteamiento del problema

Dentro del campo entomológico existe software específico que se utiliza para poder realizar la clasificación de especies de forma automatizada y entregar resultados de forma más rápida, sin embargo, existen varias restricciones que nos dificultan el uso de los mismos algunas de estas restricciones las principales dificultades son: son la versión y el tipo del sistema operativo, los entornos de desarrollo, las interfaces de usuario y la compatibilidad entre la versión el software y el sistema operativo.

Debido a las necesidades y evolución de sistemas operativos se requiere que el software a desarrollar sea compatible con las versiones más actuales, se quiere desarrollar un software que sea compatible y que tenga mejoras de experiencia de usuario como son nuevas interfaces, presentación de datos, manejo de archivos y algoritmos más rápidos que nos permitan presentar mejor la información y mejorar el rendimiento.

1.2 Justificación

El conocimiento existente acerca de un objeto, persona o cosa dan identidad a los propósitos de un proyecto. Las ciencias necesitan de otras disciplinas auxiliares para poder cumplir su propósito y así poder concluir con el trabajo que le ha sido encomendado.

La entomología es la ciencia que se encarga del estudio de los insectos, y aunque tiene muchos años de ventaja sobre la computación, depende en gran parte de las tecnologías para poder realizar todos sus procesos de forma automatizada y así poder disminuir tiempo, trabajo y esfuerzo.

Al estudiar los insectos podemos realizar diferentes procesos como identificar su anatomía, su medio ambiente, su habilidad y su clasificación. La clasificación de especies se realiza mediante diferentes técnicas que presentan puntos a favor y puntos en contra.

Una de las técnicas que presentan mayores beneficios y resultados para los investigadores es comparar especies mediante la cladística. Esta es la técnica que se utilizara para poder catalogar, identificar las familias, antepasados y los insectos que tienen las características similares.

Desarrollar software necesita como pilar comprender en los requerimientos del usuario para entregar los resultados. Una vez que se tienen los requerimientos establecidos podemos proceder a elegir las herramientas tecnológicas y metodológicas mediante las cuales se hará todo el proceso de desarrollo.

Como herramienta de desarrollo se optó por trabajar con un lenguaje de programación orientado a objetos. El lenguaje de programación Java (Oracle, 2015) es uno de los lenguajes orientados a objetos que permiten que sus aplicaciones sean instaladas en diferentes sistemas operativos además de contar con portabilidad en miles de dispositivos.

Dentro de las necesidades que se presentan existen algunas que podemos mencionar como: trabajar con archivos que contengan información y los datos sobre los cuales se harán los procesos, lectura y escritura de datos, presentación de resultados a través de diferentes formatos de archivos y exportación de reportes por correo electrónico, y estas serán satisfechas mediante la aplicación que se desarrollará.

Por lo cual, la relación entre los sistemas de información y la entomología no están fuera de lugar y esto nos da pie a que nuestra propuesta de desarrollo sea viable. Tras haber comprendido las necesidades de aplicar nuestros conocimientos para desarrollar un Sistema de Información como el que se ha descrito estamos solucionando un problema real que podrá beneficiar a muchas personas dedicadas a la de investigación entomológica.

Debido a las necesidades de clasificar, se quiere desarrollar un software que permita catalogar insectos por medio de un front-end amigable al usuario y un back-end con un algoritmo inteligente que pueda ser validado por los expertos del área de la clasificación.

1.3 Objetivo General

Desarrollar un sistema de información (software) de código abierto, independiente de la plataforma y con un ambiente amigable para el usuario con la finalidad de que este entienda y use las herramientas que el software pondrá a su disposición para la generación de árboles filogenéticos.

Se utilizará como herramienta metodológica la Ingeniería de Software (se realizarán prototipos en forma evolutiva para una revisión continua con el especialista, por lo que la metodología que más se acerca es la espiral) y como herramientas tecnológicas el lenguaje de programación Java.

1.4 Objetivos específicos

La sistemática se ha convertido en un área de gran importancia para la Biología moderna. En los estudios de biodiversidad, cada vez es más común encontrar cladogramas como mecanismo de deducción o comparación de hipótesis sobre la historia de diversos atributos, funciones, o de los procesos genéticos y evolutivos. Por tal motivo, desarrollar un sistema que nos permita conocer la historia evolutiva tiene objetivos específicos como:

- Identificar los métodos generales para reconstrucción de filogenias que más se utilizan para poder interpretar la información
- Definir el algoritmo de generación de los arboles filogenéticos
- Si es posible, estudiar e implementar un algoritmo inteligente
- Analizar los resultados obtenidos para medir la veracidad de los resultados
- Desarrollar un sistema de información para la generación de árboles filogenéticos.

1.5 Supuesto

La integración de una plataforma de software libre y los algoritmos de reconstrucción filogenética basados en caracteres homólogos que procesen la información recabada de las investigaciones de un experto usando algoritmos heurísticos facilitará la toma de decisiones y mejorará la representación gráfica de los resultados para una mejor interpretación.

2. Trabajos Previos y Antecedentes

En el presente capítulo se comentan los principios y bases fundamentales de la entomología. Además, se comentan las diferentes ramas en las que se divide la Sistemática a fin de poder comprender el funcionamiento de los componentes que nos ayudarán a desarrollar un sistema de información para la generación de filogenias.

2.1 Definición de Entomología

La Entomología proviene del griego *éntomos* insecto, y *logos* tratado (ciencia: es el estudio científico de los insectos). De cerca de los 1,3 millones de especies descritas, los insectos constituyen más de los dos tercios de todos los seres vivos conocidos y además tienen una larga historia fósil. Para poder estudiar esta gran cantidad de especies es necesario que se auxilie de otras ramas para poder realizar un estudio específico de alguna especie.

Todos los organismos que se mueven sobre la tierra son resultado de la evolución o filogenética. Si analizamos la historia de la evolución encontraremos que se conecta a través de ancestros compartidos u otros linajes de organismos. Toda esa vida está conectada en un árbol filogenético y es uno de los mayores descubrimientos de los pasados 200 años. La rama

de la Biología que reconstruye este árbol y descubre los antepasados que brinda la distribución y la diversidad de la vida es llamada Sistemática.

Entonces, la sistemática no es más que el entendimiento de la historia de la vida. Debido a la importancia de esta rama, la Sistemática forma las bases de otras ramas de comparación en la Biología.

La palabra sistemática proviene de la raíz griega *system* que significa arreglo ordenado de cosas y se ha aplicado a los sistemas de clasificación, principalmente a aquél desarrollado por el naturalista Linneo llamado System Nature (Linnaeus, 1735). El científico Simpson define Sistemática como "*el estudio científico de las clases y diversidad de organismos y de las relaciones entre ellos*" (Ramos, 2007). En lo general es posible encontrar otros términos que refieran a la misma disciplina, como biosistemática o neo sistemática; sin embargo, no dejan de ser variantes del mismo término. (Napolés, 1990)

La Sistemática ha sido definida por algunos investigadores como uno de los puntos focales para la Biología, para otros indican que la Sistemática es la piedra angular para toda investigación biológica debido a que el punto inicial o el punto de partida siempre es un organismo y para que se pueda llevar a cabo y describir propiamente un experimento se debe conocer con seguridad el organismo que se está empleando.

2.2 Orígenes de la Sistemática

Por siglos y por diversos propósitos, naturalistas, filósofos, químicos, botánicos y zoólogos, entre otros estudiosos de la naturaleza, intentaron ordenar de alguna forma a los seres vivos. Dentro de esos intentos Aristóteles (322-384) desempeñó un papel crucial debido a que trató de establecer criterios que permiten clasificar animales y plantas de forma sistemática y jerárquica. (Goyenechea, 2006).

Las ideas de Aristóteles plantearon un pensamiento científico durante varios siglos y abrió camino seguido por muchos científicos, entre ellos el sucesor inmediato de Aristóteles en Atenas. Teofrasto (287-372 a. C) quien propuso la primera clasificación jerárquica de las plantas. Para lograr esta clasificación tomó en cuenta el sistema reproductivo, el tipo de inflorescencia, y en las de reproducción sexual el número de cotiledones (número de hojas del embrión).

Pedanio Dioscórides (40-90 d. C.) médico, farmacólogo y botánico griego analizó el valor farmacológico de plantas y animales en 5 volúmenes de su tratado de materia médica. Este tratado constituyó la principal referencia de la farmacopea de la edad media y el renacimiento.

Ese tratado, al igual que otras ideas sobre el mundo viviente, prevaleció durante muchos siglos e influyó poderosamente.

El médico y botánico bávaro Leonhart Fuchs (1501-1566) elaboró una guía de plantas con nombres comunes y descripciones morfológicas que incluían, además de aplicaciones terapéuticas, un glosario de Botánica. Basó su ordenamiento en varias características de los órganos vegetativos.

El médico y filósofo toscano Andrea Cesalpino (1519-1603) es considerado como el primer botánico en sentido moderno. Cambió el enfoque de la clasificación de las plantas, pues hizo a un lado basarla en sus aplicaciones terapéuticas y retomó el criterio de apoyarla en las características morfológicas observables de sus frutos y semillas, a esto se le conoce como fenotipo, es decir, el resultado de la interacción de la constitución genética o genotipo y ambiente.

Para finales del siglo XVII, el naturalista inglés John Ray (1627-1705) avanzó en la dirección de la descripción empírica la misma que había tomado Celsapino en oposición a la definición de órdenes racionales. Más tarde definió una especie como un grupo de individuos con ciertas características en común que se perpetúan en la progenie.

En el siglo XVIII, Georges-Louis Leclerc, conde de Buffon (1707-1788), también adoptó esa noción, llamada de aislamiento reproductivo, para definir una especie de esta forma John Ray también consideró que los sistemas de clasificación de las plantas tenían que ser naturales, para lo cual debían basarse en el mayor número posible de rasgos.

En cambio, su contemporáneo sueco Carl Linnaeus (1707-1778), catedrático de la Universidad de Upsala, como naturista botánico y zoólogo, propuso un sistema artificial de clasificación de dos nombres compuestos por género y especie. Además, agrupó los géneros en familias, las familias en clases y las clases en reinos, categorías que con el tiempo se incrementaron. Al advertir que su idea original, sólo basada en características de las estructuras reproductivas, tenía limitaciones, recurrió en adición a otros rasgos.

Por su parte, y en concordancia con las ideas de Ray, el médico y botánico francés Antoine-Laurent de Jussieu (1748-1836) ideó un método analítico de clasificación natural, basado en la continuidad de muchos caracteres morfológicos y la subordinación entre ellos.

También en el siglo XVIII, Erasmus Darwin (1731-1802), abuelo de Charles Darwin, relacionó la variación morfológica de las plantas con su modo de reproducción, que puede ser sexual, por semillas, o asexual, por estructuras vegetativas como tubérculos, gajos, raíces gemíferas u otras. Charles Darwin (1809-1882) retomó las ideas de su abuelo, por lo que describió y clasificó gran parte de los grupos entonces conocidos de plantas y animales.

Gracias a todos los científicos que aportaron sus ideas la clasificación de los seres vivos ha tenido gran desarrollo. En la figura 1 se muestra una recta de tiempo de lo descrito.

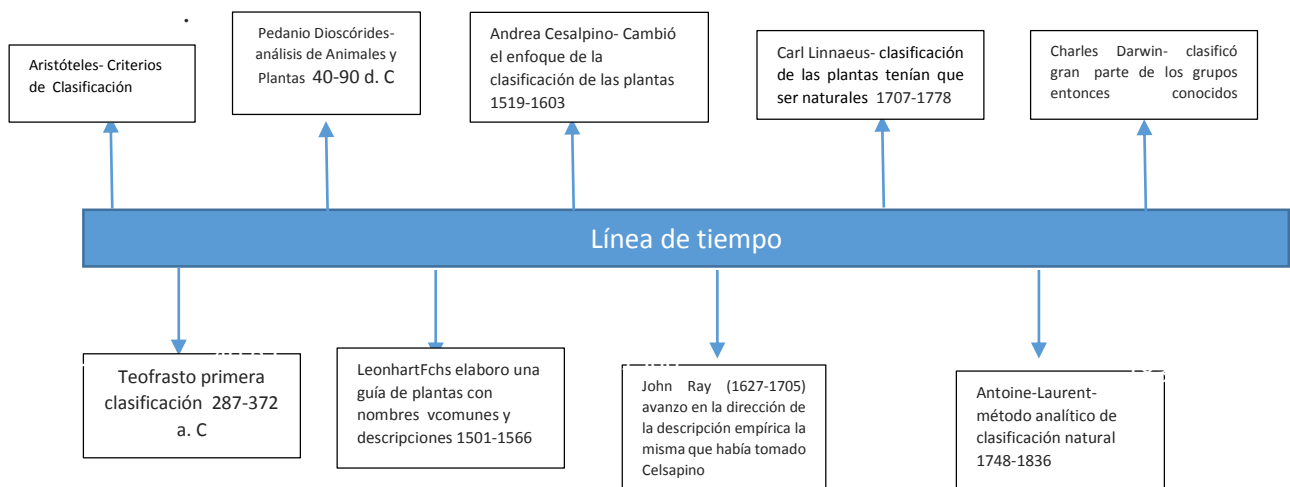


Figura 1 Línea de Tiempo de científicos que han participado en el progreso de la evolución

2.3 Taxonomía y Sistemática

La **Taxonomía** se ha definido como una forma de organizar la información biológica con arreglo a diferentes métodos (como la fenética, la cladística, la taxonomía evolutiva, criterios de tipo ecológico, paleontológico, etc). Es una disciplina eminentemente empírica y descriptiva, acumula fenómenos, hechos, objetos, y a partir de dicha acumulación genera las primeras hipótesis explicativas.

La **Sistemática** es una disciplina de síntesis, de abstracción de conceptos, de enunciado de teorías explicativas de los fenómenos observados. Por lo tanto, tiene en sí, un trasfondo teórico que supera al de la taxonomía y una vocación predictiva.

Es necesario decir que la sistemática y taxonomía son ramas de la Biología responsables de la categorización jerárquica.

La sistemática se encarga de crear sistemas de clasificación en los cuales se toma en cuenta:

- Los rasgos de similitud
- Diferencias
- Origen
- Relaciones evolutivas de cada especie

Los sistemas de clasificación se representan en forma de árbol ramificado, en cuya base se identifica al ancestro y en las ramas la descendencia de las especies que contiene.

La taxonomía se encarga de poner las reglas y procedimientos para identificar, nombrar y clasificar a cada una de las especies en las categorías o niveles de forma jerárquica, siguiendo los patrones de la sistemática.

2.4 La Filogenia como base de la Sistemática

Los científicos estudian un grupo de organismos, seleccionando características que se creen importantes y delimitan especies y grupos de especies basados en estas características. Desafortunadamente los desacuerdos siempre aparecen cuando los científicos piensan en características importantes debido a que lo que para uno es importante para otro pueda ser banal.

Por tal razón es difícil evaluar el significado de evolución de grupos clasificados por intuición porque en realidad no se sabe porque ellos representan algo real en la naturaleza o porque esos grupos no han sido definidos del todo o fueron definidos en base a un significado evolutivo como grupos artificiales.

2.5 Escuelas de la Sistemática

Cada investigador puede tomar una línea diferente para poder hacer un estudio o investigación sobre un grupo determinado de estudio, y como se mencionó con anterioridad, gracias al estudio de la evolución se desarrollaron metodologías diferentes para poder encontrar un resultado final.

Estas metodologías, pueden ser llamadas también escuelas y son las más utilizadas a la hora de la búsqueda de un árbol filogenético ya que han presentado grandes resultados en proyectos de investigación.

La Sistemática se divide en.

- Sistemática evolutiva o tradicional.
- Sistemática numérica o Fenética.
- Sistemática filogenética o cladística.

Cada una presenta ventajas y desventajas que serán descritas a continuación.

2.5.1 Sistemática Evolutiva o tradicional

Muchos de los científicos seguidores de la Sistemática evolutiva o sistemática tradicional afirman que esta escuela proviene de la herencia darwiniana y definen que este método de análisis es el mejor debido a que en ellas se refleja tanto la evolución, así como los grados de similitud y divergencia y el parentesco filogenético. Sin embargo, esta escuela no tiene un método definido con el cual se puedan contrastar las hipótesis filogenéticas diferentes, y en el caso de querer decidir cuál de las dos hipótesis es superior siempre se considera más adecuada la propuesta hecha por un taxónomo reconocido experto en el grupo en cuestión. Además, al agregar más aspectos evolutivos dentro de la clasificación que se está utilizando, el resultado es ambiguo debido a que pueden reconocer grupos no naturales o grados.

Los evolucionistas utilizan las relaciones jerárquicas en diagramas llamados filogramas. Se cree que son muy informativos porque en ellos se puede observar el tiempo en el que aparecieron, cuando tuvieron evoluciones, la extinción de algún grupo, etc.

Un grado puede ser definido con el término de evolución. La evolución es reconocida por un conjunto de caracteres adaptativos o también conocidos de adaptación con el ambiente. Muchas especies descendientes de un taxón ancestral pueden mantener las características originales de su hábitat, pero a lo largo de la evolución del grupo puede ocurrir que alguno de sus descendientes invada un nuevo nicho o hábitat lo que formaría nuevos grupos diferentes con los ancestros y los descendientes. Los nuevos grupos ligados a un taxón ancestral forman un nuevo grupo o grado evolutivo.

Los grados entonces se consideran como el criterio para la construcción de jerarquías de taxones dentro de las clasificaciones evolutivas. (Goyenechea, 2006).

2.5.2 Numérica o Fenética

Los científicos de la Fenética sugirieron un método debido a la falta de algún método robusto dentro de la sistemática evolutiva. La idea principal de la Fenética es tomar en cuenta el mayor número de características que sea posible de medir, contar y observar en los organismos y analizarlos con técnicas numéricas.

La Fenética es un intento de idear un método empírico para determinar las relaciones taxonómicas. En la práctica la Fenética no es mejor que los métodos tradicionales porque varios algoritmos se concentran en proyectar la similitud entre los organismos en cuestión. Los organismos con características similares se agrupan juntos ignorando los resultados de paralelismo o convergencia de evolución y por tanto dan lugar a grupos artificiales.

Los seguidores de la Fenética intentan obtener la mayor cantidad posible de información y son rigurosos en cuanto al análisis numérico, pero piensan que la clasificación debe hacerse con base en la similitud total más que en la genealogía. Esto nos da como resultado que no están interesados en saber cómo han evolucionado las especies ni como están emparentadas una con la otra. (Goyenechea, 2006).

2.5.3 Filogenética o cladística

La filogenia se define como la historia o crónica evolutiva de las especies. En principio no establece grupos taxonómicos como familias, géneros, etc. Su misión es conocer las relaciones evolutivas entre los grupos de especies y hay un acuerdo generalizado en que es el criterio a seguir en el establecimiento de la organización natural.

Desde que la teoría de la evolución ha ganado terreno, la sistemática ha buscado la historia de la evolución de los organismos y la ha tratado de representar en una estructura jerárquica.

La filogenética pretende descubrir las relaciones ancestrales comunes que vienen al compartir características derivadas, las relaciones se muestran en forma de árbol filogenético que nos sirve para reconstruir la genealogía de una relación genealógica.

Una de las mayores cualidades de la filogenética es que los resultados son transparentes significando que las decisiones, malas o buenas, se basan en datos que pueden ser examinados por cualquier persona que trata de entender los datos de la naturaleza, este método no depende de algún conocimiento acerca de organismos que solamente un experto puede comprender.

El concepto completo de filogenética es el uso derivado de características apomórficas para reconstruir relaciones ancestrales comunes y el agrupamiento de taxa basado en relaciones ancestrales comunes.

Una característica apomórfica es el término que se utiliza para distinguir el organismo o taxón de otros que comparten el mismo antepasado.

Independientemente del método usado para estudiar la filogenia, ésta es única. No existe más que un árbol de la vida, que comienza con el primer ser vivo sobre la Tierra y termina con todas las especies de organismos que existen en la actualidad. Será pues, trabajo del investigador de la filogenia el descubrir las relaciones evolutivas entre las especies.

Por esta razón en la actualidad se considera al Cladismo casi como la única forma de estudiar con criterios científicos estas relaciones, aunque se ha recordado que no es el único.

La necesidad de la filogenia en la clasificación es clara ya que las categorías clasificatorias dejan de ser abstracciones ideales más o menos arbitrarias para convertirse en entidades reales que expresan la perspectiva histórica única e irreplicable del mundo orgánico. De esta forma se consigue un valor predictivo en los grupos formados y, además, es refutable con la aportación de nuevas evidencias filogenéticas.

La cladística o sistemática filogenética es el paradigma actual de la taxonomía. La cladística considera que la clasificación natural es aquella basada en las relaciones genealógicas de los organismos, comúnmente expresadas en un cladogramas. Este método se ha convertido en el más utilizado por los sistemáticos de todo el mundo para la reconstrucción filogenética de grupos biológicos. Este método es el método más robusto y ha empezado a permear en otras disciplinas de la biología, la geografía, ecología, evolución, etc.

La sistemática filogenética fue propuesta por Willi Hennig en 1950 y hace uso de caracteres homólogos para reconocer grupos monofiléticos usando un método robusto y repetible, sin importar que tanto renombre tenga el taxónomo que proponga la clasificación. (Goyenechea, Sistemática: su historia, sus métodos y sus aplicaciones en las serpientes, 2006)

A continuación, se presenta una tabla comparativa que resalta las características básicas que se pueden considerar dentro de los métodos en la sistemática (*ver tabla 1*).

Tabla 1 Comparación de métodos en la Sistemática

ATRIBUTOS	FENÉTICA	CLADÍSTICA	SISTEMÁTICA ORTODOXA
Muestra las relaciones mediante un árbol o clasificación	Semejanza o desemejanza global	Genealogía	Genealogía + semejanza global
Semejanza evolutiva	Usados todos los tipos	Solamente apomorfías	Solamente homologías
Peso de los caracteres	No usado	Generalmente no usado	Usado
Homología	No considerada	De importancia capital	Importante
Fósiles	No usados	Pueden ser considerados, pero no de más importancia que las especies vivas	Puede ser muy importante

Datos ecológicos y evolutivos	No usados	Raramente usados	Puede ser muy importante
Tasas de evolución	No considerable	No considerable	Muy importantes
Transformación del árbol en una clasificación	Sin reglas generales; para delimitar los taxones se escogen niveles arbitrarios de semejanza	La clasificación muestra precisamente modelos de ramificación adyacentes	La clasificación refleja tanto modelos de ramificación como grados de diferencia entre taxones

Resumiendo, en la Fenética numérica se agrupan las especies de acuerdo a sus características externas. Representa un intento de clasificar los organismos basándose en su similitud global, normalmente establecida sobre caracteres morfológicos u otros rasgos notables, sin tener en cuenta su filogenia.

Mediante la Fenética se persigue encontrar un modelo de ordenación de las características de los organismos que reduzca los patrones de variación que con frecuencia son muy amplios a unos márgenes manejables convirtiéndolos en valores numéricos. En la práctica esto se hace midiendo series de variables, a menudo basadas en características morfológicas, que se ordenan, mediante tratamiento matemático, en gráficos multidimensionales.

El problema de este método consiste en que para hacer comprensibles los resultados hay que renunciar a parte de la información decidiendo, de una forma un tanto arbitraria, cuáles son los caracteres de mayor peso, es decir los que más influyen en el resultado final, y cuáles son los menos relevantes y pueden menospreciarse.

En cambio, la cladística analiza matrices de datos para producir un diagrama de árbol denominado cladograma.

Un cladograma se puede definir como una hipótesis grafica de relaciones genealógicas entre especies (conocidos como taxones).

La robustez de estas hipótesis se puede verificar mediante medidas de bondad de ajuste de la matriz a cladogramas como el índice de consistencia o técnicas de remuestreo que ponen a prueba la robustez de cada grupo.

El método cladista ha experimentado cambios desde su concepción original, sin embargo, la esencia del método se ha mantenido. Los biólogos ahora cuentan con más herramientas para evaluar sus hipótesis filogenéticas.

Ahora bien, la razón de la factibilidad y viabilidad del método cladista descansa sobre dos propiedades de los seres vivos que podemos identificar a continuación:

- En primer lugar, los seres vivos forman linajes. El científico Simpson en el año 600 definió a un linaje como un conjunto de organismos, interconectados a través del tiempo y del espacio, por la transferencia de material genético, de los progenitores a su descendencia. Un ejemplo sencillo de linaje es la agrupación hijo-

padre-abuelo, tres organismos relacionados entre sí por la transferencia de información genética.

- En segundo lugar, los progenitores producen descendientes con modificaciones heredables, que se integran a la variabilidad de los linajes. Si el linaje se divide, la variación es la base de la que surgen nuevos linajes. (Lara, 2015)

Debido a estas características se ha determinado el uso de la Cladística como la corriente de investigación sobre la cual podremos basar los resultados. Las herramientas que ayudan a comprobar la robustez son de gran utilidad para poder sostener resultados y encontrar pruebas que puedan ser aceptadas.

3. Principios Básicos

Para los años 450 A. C. el Filósofo Sócrates, en su búsqueda interminable sobre el conocimiento, hizo mención de uno de los dilemas más grandes que ha tenido el ser humano “la investigación”.

Sócrates sostenía que “la investigación es el objetivo primordial y el fin básico de la existencia del ser humano” (Bastar, 2012), y decía esto debido a que como seres pensantes tenemos la necesidad de saber cómo funciona el mundo a nuestro alrededor, entender problemas matemáticos muy complejos o simplemente concebir los hechos que suceden día con día.

Para comprender todo lo relacionado a un mini mundo o universo en discusión es necesario que entendamos como funciona cada elemento que está contenido en ese universo, de tal forma que de lo particular caminemos hacia el funcionamiento total y comprendamos como es que nuestros elementos interactúan entre sí para dar vida a un todo.

Dentro de la Sistemática Filogenética o Cladística, como es también conocida, existen un sinnúmero de elementos que participaran dentro de un análisis filogenético y que por ende nos basaremos en ellos para construir modelos gráficos de representación (Cladogramas) que nos muestren los resultados de los análisis que hemos hecho.

La importancia de integrar todos los elementos dentro la cladística será el éxito de nuestro análisis de tal forma que los resultados se acercaran lo más posible al resultado deseado por el usuario.

3.1 Elementos Dentro de la Cladística

La cladística o sistemática filogenética estudia la diversidad orgánica a través del reconocimiento de las relaciones genealógicas de los organismos. Por consiguiente, los métodos de cladística agrupan organismos que comparten determinadas características.

Estas características provocan que los taxos (grupos) similares sean agrupados y aquellos que no comparten muchas similitudes sean declarados no aptos para entrar dentro de este agrupamiento.

Las relaciones mencionadas son representadas mediante una estructura jerárquica que representa a los grupos que fueron agrupados. Habitualmente estas representaciones gráficas son conocidas dentro de la cladística como arboles filogenéticos o cladogramas los cuales son denominados así por las ramas que los conforman y los nodos finales que lo componen.

Antes de poder comenzar a crear los cladogramas es necesario que establezcamos algunos principios fundamentales de la sistemática filogenética.

Entre los conceptos más importantes que se usan en cladística están los términos como la parsimonia y el cladograma que son los elementos fundamentales sobre los cuales se basan una serie de conceptos y definiciones que serán descritas en este capítulo.

Parsimonia. Es el principio que dice que la naturaleza se comporta de manera sencilla, es decir, que no busca caminos intrincados, por lo que al realizar un análisis cladístico en donde se obtienen varios cladogramas como resultados, se prefiere el que sea más parsimonioso, es decir, que tenga el menor número de pasos.

Cladograma. Es el diagrama de ramificación resultado de un análisis cladístico y en él se reflejan las relaciones filogenéticas de los taxones terminales. Según (Morrone, 2000) se pueden nombrar las partes del cladograma como ramas a cada una de sus divisiones, donde en la punta se colocan los taxones terminales.

Otros componentes importantes son los nodos, los cuales son las intersecciones de dos ramas, y denota que existe una relación desde el nodo y a todas las ramas que parten de él. Hay dos tipos de nodos, el nodo basal o raíz que representa la intersección o punto de partida del cladograma y los nodos internos que son las intersecciones que se dan entre ramas intermedias del cladograma (Goyenechea, 2006).

3.2 Las funciones de la Sistemática

La sistemática, como parte de las ciencias biológicas, tiene funciones importantes las cuales son:

- Proveer, mediante la clasificación, el marco conceptual a través del cual los biólogos pueden comunicar información acerca de los seres vivos.
- Proporcionar, mediante cladogramas, las bases para proveer diferentes interpretaciones evolutivas.
- Predecir, mediante cladogramas y las clasificaciones derivadas de los mismos, propiedades de los organismos recién descubiertos y poco conocidos. (Morrone, 2000).

Estas funciones expresan lo que en realidad se tiene que hacer. Por eso, para llevar a cabo un estudio sistemático aplicando la metodología cladística básicamente se siguen cuatro pasos:

1. Seleccionar los taxones que serán las unidades de nuestro estudio.
2. Seleccionar los caracteres que brindarán la evidencia sobre las relaciones genealógicas de los taxones estudiados.

3. Descubrir las relaciones genealógicas de los taxones analizados y expresarlas en un cladograma.

4. Traducir las relaciones genealógicas del cladograma en una clasificación formal (Morrone, 2000).

Un paso muy importante para realizar un análisis cladístico consiste en seleccionar los caracteres que permitirán evidenciar las relaciones filogenéticas de los taxones estudiados.

La tarea de seleccionar taxones, que serán nuestras unidades de estudio, es lo más básico y esencial de nuestro análisis; dentro de esta etapa es necesario que se determinen los taxones que participaran y los caracteres que se utilizaran para poder generar el análisis respectivo. En este punto cabe destacar que el listado de taxones y la matriz de características son definidas por el experto que pretende generar el cladograma respectivo. Es necesario también determinar también cuales caracteres son primitivos y cuales son derivados y por consiguiente es ineludible dar a conocer algunas definiciones antes de comenzar a explicar el resto de la información.

3.3 Caracteres

Un carácter es cualquier atributo que podemos observar en un organismo cuyas diferentes manifestaciones se denominan estados. Para definir el valor de un estado respectivo a un carácter específico se determina por medio del número de elementos que tiene el atributo.

Supongamos el número de patas, los valores pueden ser: 2 patas, 4 patas. Para el carácter número de alas los estados respectivos podrían ser 2 alas, 4 alas, etc. O al definir el número de manchas que son observables en algún organismo nuestros estados que podría tomar el atributo podrían ser de 2 manchas, 4 manchas, etc.

Existen diferentes formas para referirse a estos estados en cladística.

- **Estado Plesiomórfico:** Es aquel estado que es presentado en principio de la relación, se podría decir que es el primero en surgir en el tiempo, se infiere que se haya o que se hallaba en el ancestro del grupo de estudio.
- **Estado Apomórfico:** Es aquel que surge a partir de un estado Plesiomórfico
- **Simplesiomorfía:** Estado plesiomórfico que está presente en dos o más taxones
- **Autapomorfía:** es un estado apomórfico presente en un único taxón
- **Sinapomorfía:** Es un estado apomórfico compartido por dos o más taxones

Todos estos términos son relativos, ya que cuando un estado surge en una especie es una Autapomorfía (Figura 2a), si luego se produce un evento de especiación se convierte en sinapomorfía (Figura 2b) y si más tarde, con otro evento de especiación, cambia a otro estado

diferente (otra Autapomorfía), el estado original de todo el grupo pasa a ser plesiomórfico (Figura 2c).

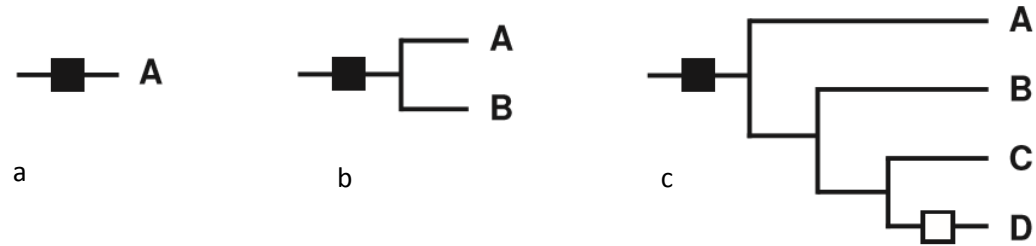


Figura 2 (a) Autapomorfía; (b) sinapomorfías; (c) simplesiomorfía

- **Homoplasia.** En los casos en que un carácter se desarrolla independientemente a partir de ancestros diferentes hablamos de homoplasia.
- Según provengan del mismo o de diferentes estados, se consideran paralelismos o convergencias, respectivamente (Fig. 3a, b). También puede ocurrir que una de las sinapomorfías de un grupo se pierda en uno de los descendientes, que entonces posee el estado plesiomórfico. En este caso hablamos de reversiones (Fig. 3c) (Morrone, 2000).

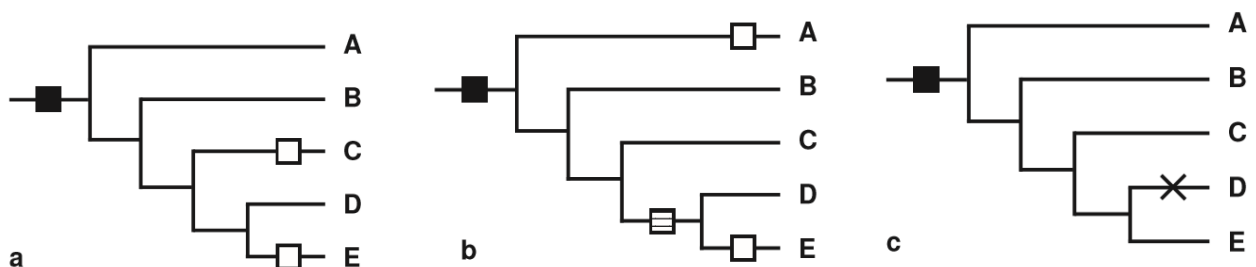


Figura 3 (a) Paralelismo; (b) convergencia; (c) reversión

3.4 Matrices de Datos

Una vez que se el experto ha obtenido el conjunto de datos a analizar es necesario que se plasmen en una estructura de datos que pueda representar plenamente los taxos que se tienen y cada característica que será tomada en cuenta para el análisis. Para esto, es común que encontremos los datos representados mediante una matriz en la cual los taxones suelen representarse en filas y los caracteres en columnas.

Es de suma importancia que conozcamos el término de grupo externo también llamado out group. La función de este grupo es que lo podamos utilizar como un grupo de comparación sobre el cual nos basaremos para poder hacer los cálculos necesarios y además poder enraizar los cladogramas. Sino los incluimos, el cladograma carecerá de raíz. Más adelante se dará a conocer la importancia de los cladogramas con raíz y sin raíz.

Un ejemplo de esta representación es presentado en la Tabla 2:

Tabla 2 Matriz de datos ejemplo

	1	2	3	4
Taxa 1	0	0	1	0
Taxa 2	1	0	0	1
Taxa 3	1	1	0	0
Taxa 4	1	1	0	1

Una matriz con información real puede ser nombrada como se muestra en la Tabla 3

Tabla 3 Ejemplo de conjunto de taxas

1	Pachymerus
2	Kythorinus
3	Megacerus
4	Callosobruchus
5	Conicobruchus
6	Rhipibruchus
7	Pectinibruchus
8	D. atrolineatus
9	D. walker
10	D. lunae

La tabla 4 muestra las características a representar.

Tabla 4 Ejemplo una Matriz de Datos Completa

1	Pachymerus	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
2	Kythorinus	1	1	1	1	0	1	0	0	1	1	0	0	1	1	2	1	1	0	0	1
3	Megacerus	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	0	1	0	1	2
4	Callosobruchus	1	1	1	1	1	1	0	1	0	1	0	1	0	2	1	0	1	1	2	1
5	Conicobruchus	1	1	1	1	1	1	1	0	0	1	0	0	0	1	1	1	0	1	2	1
6	Rhipibruchus	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	1	1
7	Pectinibruchus	1	1	1	1	1	1	0	1	1	1	0	0	0	1	1	1	1	1	1	1
8	atrolineatus	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	2	1
9	walker	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	2	1
10	lunae	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	2	1

3.5 Cladogramas

Un cladograma (Fig. 4) es un dendrograma que refleja las relaciones genealógicas de los taxones terminales. Hay varios términos empleados para describir las partes de un cladograma

- **Raíz o nodo basal:** Es la base o punto de partida del cladograma.
 - **Nodos internos o componentes:** Son los puntos de ramificación del cladograma, es decir que están conectados con dos o más nodos o taxones terminales.
 - **Ramas internas o internodos:** Son los segmentos que unen nodos internos entre sí.
 - **Ramas terminales.** Son los segmentos que unen nodos internos y taxones terminales.
 - **Taxones terminales:** Son las unidades en estudio, es decir los taxones que se hallan situados en los extremos de las ramas terminales y están conectados con un solo nodo interno o con la raíz.

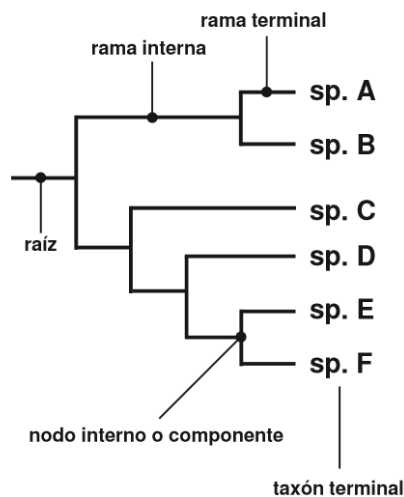


Figura 4 Partes de un Cladograma

Hay varios términos que se emplean para describir los distintos tipos de cladogramas

- Cladogramas dicotómicos, binarios o totalmente resueltos. Son los cladogramas en que ningún nodo interno se conecta con más de dos nodos o taxones terminales (Fig. 5).

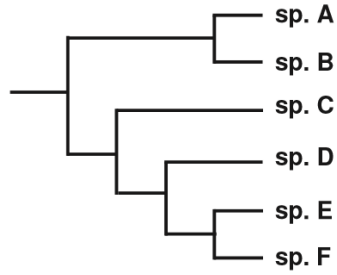


Figura 5 Cladogramas dicotómicos

- Cladogramas politómicos o parcialmente resueltos. Son los cladogramas que contienen uno o más nodos internos conectados con más de dos nodos internos o taxones terminales. En la figura 6, vemos una tricotomía basal que conduce a AB, C y DEF

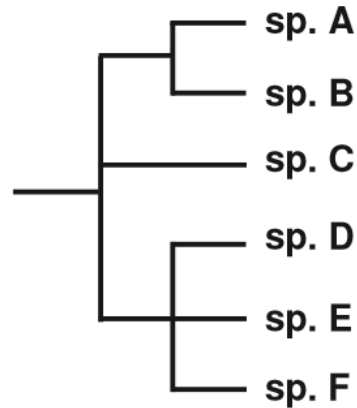


Figura 6 Cladogramas Politómicos

- Cladogramas no enraizados. Son los cladogramas en que no hay nodo basal o raíz (Fig. 7a).
- Cladogramas enraizados. Son los cladogramas que tienen un nodo basal o raíz que les imparte dirección. A partir de un cladograma no enraizado podemos obtener diferentes cladogramas enraizados, según en qué parte del mismo coloquemos la raíz (Fig. 7b, 7c, 7d, 7e, 7f).

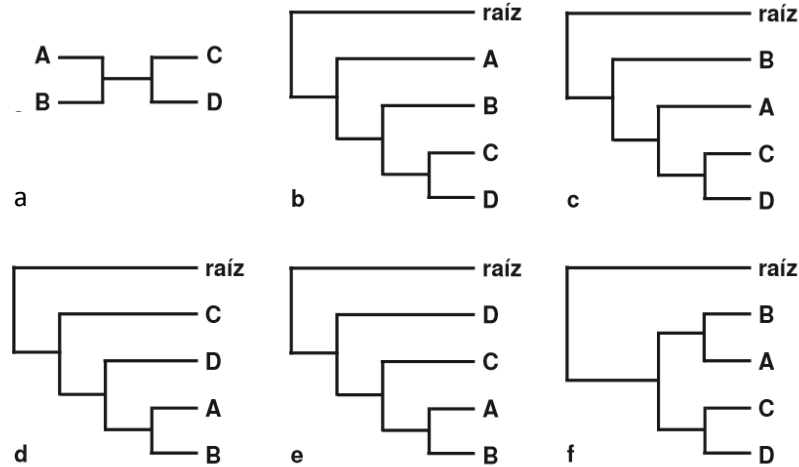


Figura 7 Cladogramas con raíz y sin raíz

3.6 Estadísticos descriptivos

Existen varios valores estadísticos que permiten determinar el grado de homoplasia de un cladograma

3.6.1 Longitud del Cladograma

La longitud del cladograma se representa por el número de pasos necesarios para sustentar las relaciones genealógicas de los taxones en el mismo cladograma. Cuando mejor sea el ajuste de los caracteres al cladograma, menor será el número de homoplasias y, por ende, menor será la longitud del cladograma (es decir, será más simple o más parsimonioso).

3.6.2 Índice de consistencia

El índice de consistencia cuantifica la homoplasia relativa de un carácter. Se calcula dividiendo el número de pasos esperados (dado el número de estados del carácter) entre el número real de pasos (Ecuación 1).

$$ci = \frac{m}{s} \quad (1)$$

Donde:

ci = índice de consistencia

m= cantidad mínima de cambios para el carácter (es igual al número de estados menos 1)

s= número real de pasos del cladograma

Cuando no hay homoplasias, m=s y ci=1, de lo contrario cuanto mayor sea la cantidad de homoplasia el valor de m será mayor y el valor de ci disminuirá.

Para calcular el valor general de la homoplasia del cladograma, podemos sumar los ci de todos los caracteres y así obtener ci general.

El índice de consistencia tiene dos inconvenientes, por una parte, se relaciona inversamente con el número de taxones y de caracteres por lo que no será útil para comparar cladogramas obtenidos a partir de distintas matrices de datos. Por otra parte, resulta sensible a los caracteres no informativos, como las Sinapomorfías de todo el grupo en estudio y las Autapomorfías. Por esta razón, es recomendable excluir del análisis a los caracteres no informativos cuando calculamos el índice de consistencia.

3.6.3 Índice de retención.

Este índice cuantifica la homoplasia observada en un carácter en función de la homoplasia posible. Se calcula mediante la siguiente Ecuación:

$$ri = \frac{(g-s)}{(g-m)} \quad (2)$$

Donde

g: mayor cantidad posible de cambios que podría tener el carácter en el cladograma

m: cantidad mínima de cambios (es igual al número de estados menos uno)

s: número real de pasos

3.6.4 Índice de consistencia rescalado

Aun cuando el ajuste de un carácter en el cladograma sea el más pobre posible, el índice de consistencia nunca llegará a ser valor cero, aunque esto si puede ocurrir con el de retención. Por esto, Farris propuso el índice de consistencia rescalado que simplemente surge de multiplicar el valor del índice de consistencia por el índice de retención (Ecuación 3).

$$rci = ci \times ri \quad (3)$$

3.6.5 Grupos monofiléticos.

También conocidos como clados o grupos naturales. Son aquellos que incluyen todos los descendientes de un ancestro común, es decir que realmente existen, como resultado de la evolución. Pueden ser reconocidos por compartir una o más sinapomorfías o por poseer una combinación particular de caracteres. En el cladograma de la figura 8, los grupos ABCDEFG, BCDEFG, CDEFG, DEFG, DE y FG son monofiléticos.

3.6.6 Grupos parafiléticos

También conocidos como grados. Son aquellos que excluyen algunos de los descendientes del ancestro común, y están basados en simplesiomorfías. En el cladograma de la figura 8, el grupo ABCDE, definido por carecer de la sinapomorfía 6, es un grupo parafilético.

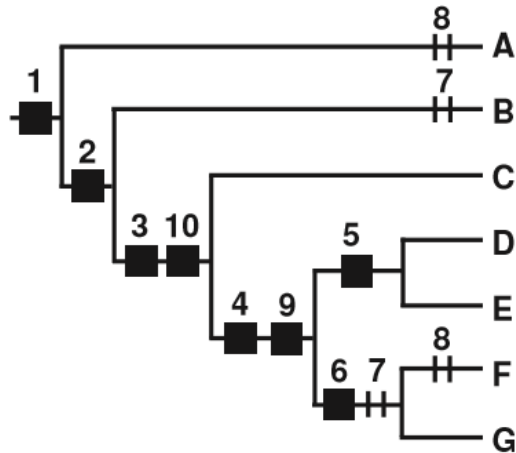


Figura 8 Cladograma con grupo Parafilético

4. Métodos de Inferencia Filogenética

De acuerdo con la información obtenida hasta ahora, hemos comprendido que el objetivo es generar a partir de cierta información un árbol, el cual podrá expresar la relación que existe entre las especies y su respectiva evolución.

Para determinar estos árboles los científicos utilizan diversos métodos que son basados en principios matemáticos sin embargo cada científico elige el método sobre el cual ha de trabajar.

Los métodos con los cuales se puede trabajar son métodos basados en matrices de distancias que previamente se tienen calculadas, es decir, la primera etapa de recolección de datos es un trabajo muy importante que solo un experto puede validar. Muchas de estas matrices de datos han sido calculadas en base a características moleculares de ácido desoxirribonucleico.

El ADN es una molécula formada por dos filamentos. Cada filamento está compuesto por una cadena que a su vez está compuesto por A (Adenina), T (timina), C (Citosina), y G (guanina). Estas cadenas se relacionan entre sí para formar una secuencia de ADN.

Las secuencias de ADN son cadenas que se relacionan entre sí (Fig. 9)

```
S = ATGACCAACATCCGAAAATCCCACCCGCTAATCAAATTATCAATCACTCATTATCG
ACCTACCAACCCCATCAAACATCTCATCTGATGAACTTTGGCTCCCTTTTAGGAATATGC
```

Figura 9 Secuencia de ADN

En base a las secuencias de ADN existe un proceso que es llamado Alineamiento de secuencias y se puede definir como comparar zonas específicas entre dos secuencias para identificar patrones para poder establecer así una función similar y un origen evolutivo entre especies.

Todos los estudios evolutivos de grupos de organismos se basan en la elección de caracteres apropiados para la reconstrucción de sus filogenias. Estos caracteres deben

cumplir dos requisitos, ser homólogos (en todos los organismos de estudio) e independientes entre sí. Los caracteres homólogos son aquellos que tienen el mismo origen y cumplen la misma función.

La naturaleza de los caracteres puede ser muy variada. Cualquier fuente de información filogenética válida y contrastada puede proporcionar caracteres fiables para un estudio evolutivo. Los principales estudios evolutivos han sido en base a caracteres morfológicos y los caracteres moleculares.

El desarrollo e implementación de algoritmos basados en caracteres moleculares como las secuencias de ADN son un marco importante para la investigación en el área biomédica, sin embargo, en este proyecto trabajaremos con caracteres morfológicos para la creación de matrices.

4.1 Teoría de Grafos

La Teoría de Grafos es una parte importante dentro de las matemáticas y el cómputo. Los problemas que se presentan en cómputo son solucionados en ocasiones mediante Grafos. En 1736 el matemático suizo Leonard Euler publicó un artículo llamado “La solución de un problema a la geometría de posición”, en este artículo Leonard Euler da la solución al problema de los 7 puentes de Königsberg.

En este problema se plantea el recorrido de una ciudad cruzando cada uno de los siete puentes exactamente uno a la vez, los habitantes de esa ciudad pensaron que no se podía, pero nadie hasta ese momento había dado algún argumento que comprobara su idea.

Euler dio una propuesta matemática a esto de tal forma que tomó solo la información importante y no tomó datos como áreas, longitud entre puentes sino concentró en la relación entre ciudades y puentes.

Un grafo $G = (V, A)$ es una colección de puntos llamados vértices V , unidos por líneas llamadas aristas A . Cada arista une dos vértices.

Las aristas son representadas mediante líneas rectas o arcos. Cuando una arista conecta el nodo a sí mismo es conocido como lazo.

En teoría de grafos es común que las aristas tengan una dirección. Puede ser que el nodo A vaya hacia el nodo B o viceversa. De tal forma que se podría denotar como $\{A, B\}$ (Fig. 10). →

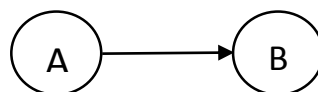


Figura 10 Grafo dirigido

Los grafos deben de cumplir con las siguientes características:

- Un sub grafo de un grafo es un subconjunto de vértices del grafo original, y un conjunto de aristas entre estos.
- El grado de un vértice v es el número de aristas que lo contienen.

- Dos vértices u, v se dicen adyacentes si existe una arista que los contiene, esto es si $\{u, v\} \in A$.
- Una trayectoria es una sucesión de vértices con la propiedad de que cada vértice es adyacente al siguiente y tal que en la correspondiente sucesión de aristas todas las aristas son distintas. Es permitido que un vértice aparezca en una trayectoria más de una vez.

4.2 Grado de un grafo

- Grado de incidencia positivo: El grado de incidencia positivo de un nodo n_j es el número de arcos que tienen como nodo inicial a n_j . Ejemplo: El grado de incidencia positivo de 1 es igual a 3 (Fig. 11).

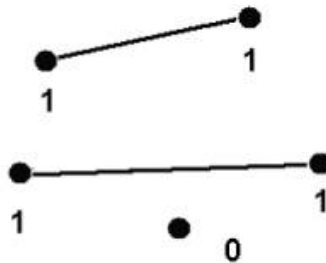


Figura 11 Grado de Incidencia Positivo

- Grado de incidencia negativo: El grado de incidencia negativo de un nodo n_j es el número de arcos que terminan en n_j . Ejemplo: El grado de incidencia negativo de 1 es igual a 1 (Fig. 12).

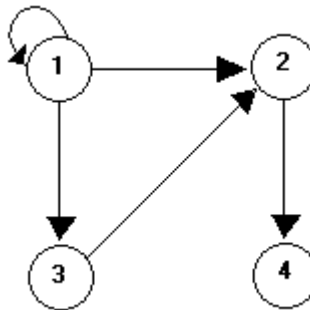


Figura 12 Grado de Incidencia Negativo

- Grado de un nodo: Para grafos es el grado de incidencia positivo menos el grado de incidencia negativo del nodo. Ejemplo: El grado de 1 es igual a $3 - 1 = 2$, el grado del nodo 4 es $2 - 2 = 0$. Para grafos no dirigidos es el número de líneas asociadas al nodo (Fig. 13).

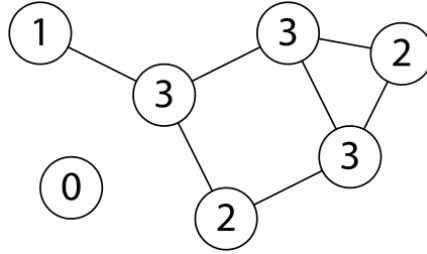


Figura 13 Grado de un nodo

En este contexto existe un tipo de grafo que es especial y es conocido como árbol. Un árbol es un grafo que tiene la cualidad de no contener ciclos, es decir es un tipo de grafo que es aciclico, pero a la vez es convexo.

Los árboles se crean en base a nodos. Un nodo es la unidad sobre la cual se puede construir el árbol y puede tener cero o más nodos hijos conectados a él. Se dice que el nodo A es padre de B si existe un enlace desde A hasta B o se puede decir también que B es hijo de A. Dentro de los árboles solo puede existir un único nodo que no tiene padre, pero es de donde salen todos los demás nodos hijos y este es conocido como el nodo raíz. Un nodo que no tiene hijos se conoce como hoja. Los demás nodos (tienen padre y uno o varios hijos) se les conoce como rama.

4.3 Construcción de Árboles Filogenéticos

Un árbol es una colección de elementos llamados nodos, uno de los cuales se distingue como raíz, junto con la relación que impone una estructura jerárquica entre los nodos **Fuente especificada no válida.** Formalmente un árbol se puede definir de manera recursiva de la siguiente forma:

- La estructura puede ser vacía
- Un conjunto finito de uno o más nodos, tal que existe un nodo especial llamado raíz, y donde los restantes nodos están separados en $n > 0$ conjuntos distintos, cada uno de los cuales es a su vez un árbol (sub arboles del nodo raíz).
- La definición implica por lo tanto que cada nodo del árbol es raíz de algún sub árbol contenido en el árbol principal. (Fig. 14).

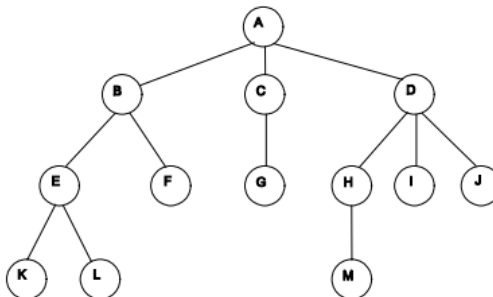


Figura 14 Árbol

Una vez que se ha determinado el grupo de especies y la matriz de distancias genéticas podemos comenzar a construir los arboles filogenéticos a través de los métodos que se van a describir más adelante, empero de todo esto es necesario decir que se pueden construir muchos árboles y cada uno de estos constituirá una hipótesis evolutiva diferente.

Una parte a considerar dentro de la Cladística es que no es un sistema intuitivo, sino que se basa en métodos empíricos de reconstrucción de filogenias usando unas reglas evolutivas estrictas como los ancestros comunes unidos a través de Sinapomorfías y en esto se basan las diferentes hipótesis evolutivas de tal forma que el experto podrá elegir la mejor de ellas.

Hemos hablado de los arboles filogenéticos en términos biológicos, pero matemáticamente un árbol filogenético puede expresarse dentro de la teoría de grafos como un árbol $H = (V, T)$ donde V es el conjunto de nodos externos (especies) e internos (ancestros) y T es el conjunto de aristas (ramas) de H .

Dado un conjunto $s = \{s_1, s_2, s_3, \dots, s_n\}$ de especies, un árbol filogenético $H = (V, T)$ es un árbol con n hojas y m nodos internos de grado 3.

Como H es un árbol entonces tiene $n+m-1$ aristas, además por técnicas de enumeración de árboles filogenéticos tenemos que H tiene $2n-3$ aristas, por tanto, el número de nodos internos es: $n-3$. Así H es un árbol con $|V| = 2n-2$ y $|T| = 2n-3$.

4.4 Raíz de un árbol filogenético

Los arboles filogenéticos pueden estar enraizados o no. En los arboles enraizados existe un nodo en particular llamado raíz a partir del cual comienza a desprenderse el camino evolutivo que se va formando. Un árbol no enraizado solamente especifica las relaciones de parentesco entre los taxones, pero no define el camino evolutivo (Fig. 15).

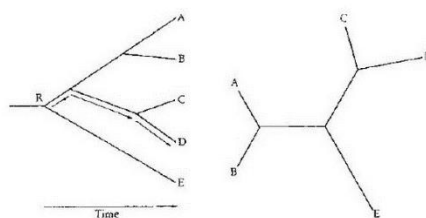


Figura 15 Árbol con raíz y sin raíz

En base al número de taxones que se utilizaran en el estudio existirá una gran diversidad de árboles. Por ejemplo, si hablamos de 3 especies A, B, C. pueden existir 3 árboles con raíz y uno sin raíz como se muestra en la figura 16.

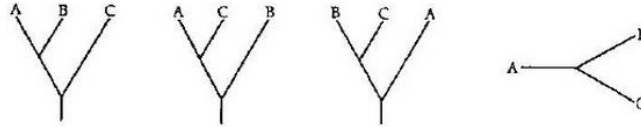


Figura 16 Árboles generados con 3 especies

4.5 Complejidad Matemática

Según (Rodríguez Catalán, 2001) el número de posibles arboles enraizados para n taxones se observa en la ecuación 4.

$$N_r = \frac{(2n-3)!}{2^{n-2}(n-2)!} \quad \text{Para } n > 2 \quad (4)$$

Donde N_r es el número de árboles con raíz y n es el número de taxones que participan en el análisis.

Y el número de árboles sin raíz se puede observar en la ecuación 5.

$$N_u = \frac{(2n-5)!}{2^{n-3}(n-3)!} \quad (5)$$

Donde N_u es el número de árboles sin raíz y n es el número de taxones que participan en el análisis.

El número de posibles arboles enraizados para n taxones es igual al de los arboles sin raíz para $n-1$ taxones. Ambos números se incrementan rápidamente a medida que n aumenta. Puesto que solo uno de esos árboles representa correctamente la verdadera relación evolutiva resulta difícil inferir la topología de un árbol cuando n es elevado como se puede observar en la figura 17.

Número de especies	Árboles con raíz	Árboles sin raíz
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10,395	945
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025
11	654,729,075	34,459,425
12	13,749,310,575	654,729,075
13	316,234,143,225	13,749,310,575
20	8.2×10^{21}	2.22×10^{20}

Figura 17 Número de árboles con y sin raíz

De este modo puede verse que a partir de 12 especies el número de árboles a cuantificar crece en forma inmensurable (En cómputo sería un problema intratable) y por

tanto se utilizan heurísticas para calcular todos los posibles árboles. Esto permite que el experto pueda analizar los árboles generados y escoger el mejor según su experiencia.

Los métodos que se establecen a continuación pueden ser clasificados en 3 grupos diferentes.

- Métodos basados en Distancias
 - WPGMA
 - Neighbor joining
- Métodos basados en caracteres (Criterio de Optimización)
 - Hennig
 - Máxima Verosimilitud (Maximum Likelihood Methods)
 - Parsimonia (Maximum Parsimony Methods)
- Métodos Bayesianos

4.6 Métodos Basados en Distancias

Los métodos basados en distancias utilizan principios matemáticos y estadísticos con el fin de unir aquellos grupos que son más parecidos entre sí. Los métodos más conocidos son:

4.6.1. WPGMA (Weighted Pair Group Method Using Arithmetic Average)

Este fue uno de los primeros métodos de agrupamiento y el más sencillo para la construcción de árboles con raíz. El WPGMA entrega como resultado un árbol en el cual las distancias entre la raíz y cada especie son las mismas (lo que se conoce como distancias ultra métricas).

Este algoritmo recibe un conjunto $S = \{1, 2, \dots, n\}$ de especies y una matriz $D \in \mathbb{R}^{n \times n}$ que contiene las distancias evolutivas entre todos los pares de especies.

Cada vez que especies son agrupadas, son reemplazados por un nuevo nodo (ancestro hipotético), el mismo que entra a formar parte del conjunto Q de nodos por procesar.

La distancia del ancestro hipotético hacia los demás nodos en Q es el promedio simple de las distancias de los nodos agrupados. El proceso se repite mientras $Q \neq \emptyset$.

4.6.2 UPGMA

Este método fue introducido como parte de los métodos de agrupamiento. Similar al método WPGMA, su diferencia radica en que las distancias son calculadas con un promedio aritmético que depende del número de especies de cada grupo. Si no se quiere trabajar con un simple promedio, el UPGMA es una buena opción.

Cada búsqueda del elemento más pequeño dentro de la matriz de distancias demanda una cantidad proporcional a n operaciones. Sin embargo, el proceso puede hacerse más eficiente creando una lista de tamaño n que almacene para cada columna, el índice de la fila

que contiene el elemento más pequeño (o viceversa), así la búsqueda del mínimo requiere una cantidad proporcional a n operaciones.

El trabajo extra requerido para mantener actualizada esta lista luego de cada agrupamiento es proporcional a n . En total esta variante del algoritmo requiere n operaciones.

La metodología de este método es similar al método WPGMA, diferente en la 2 forma de calcular las distancias.

Tabla 5 Métodos de reconstrucción filogenética, ventajas y desventajas

Métodos	Ventajas	Desventajas
Matrices de distancias	Rápidos	Secuencias son transformadas y se pierde información
Máxima parsimonia	Es robusto si las ramas son cortas	Mala representación cuando las ramas tienen una considerable longitud
Máxima verosimilitud	La verosimilitud captura toda la información de los datos bajo el modelo dado	Computacionalmente lento
Inferencia bayesiana	Las distribuciones a priori pueden ser especificadas	Dificultad para determinar un criterio de parada en las Cadenas de Markov Monte Carlo

5. Hennig

La argumentación de Hennig considera la información de cada carácter uno a la vez. Es fácil de entender mediante un pequeño ejemplo tomado de (Lipscomb, 1998). En este ejemplo tomaremos como base una matriz de datos que contiene 4 grupos de estudio (taxas) y 6 características (ver Tabla 6)

Tabla 6 Tabla de datos con 5 especies

	Características				
	1	2	3	4	5
Outgroup	0	0	0	0	0
A	1	0	0	0	1
B	1	1	0	1	0
c	1	0	1	1	0

1.- El carácter 1 une los taxas (grupos) A, B y C porque ellos comparten el carácter apomorfico 1 (Fig. 18).

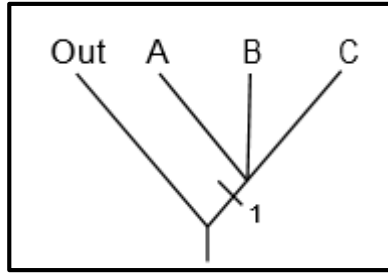


Figura 18 Árbol generado con el carácter 1 (Lipscomb, 1998)

2.- Carácter 2 – el carácter derivado es encontrado solo en el taxón B, y no provee mucha información sobre las relaciones entre taxas (Fig. 19).

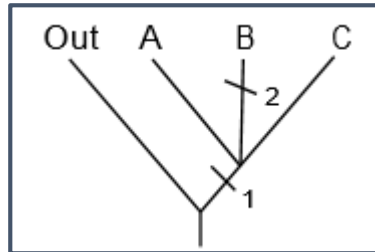


Figura 19 Árbol generado con el carácter 2 (Lipscomb, 1998)

3.- Carácter 3 el carácter derivado es autopomorifico para el grupo C (Fig. 20).

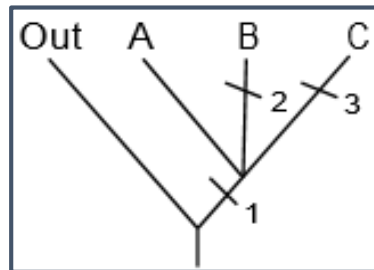


Figura 20 Árbol generado con el carácter 3 (Lipscomb, 1998)]

4.- Carácter 4 el carácter derivado es sinapomorifico y une los taxas A y B (Fig. 21)

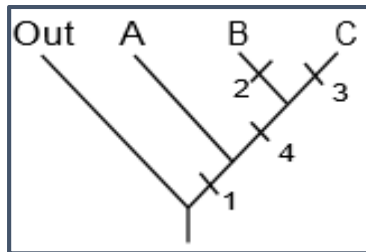


Figura 21 Árbol generado con el carácter 4 (Lipscomb, 1998)

5.- Carácter 5 El carácter derivado es un autopomorifico para el taxón A (Fig. 22)

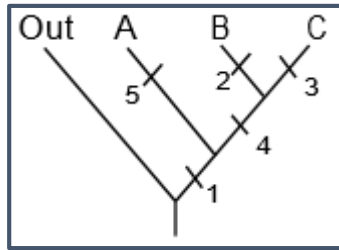


Figura 22 Árbol generado con el carácter 5 (Lipscomb, 1998)]

Las matrices de datos reales raramente son así de simples. Sin embargo, el concepto es el mismo. Dentro del siguiente diagrama de flujo se representa el funcionamiento de nuestro algoritmo (Fig. 23).

5.1 Diagrama de Flujo

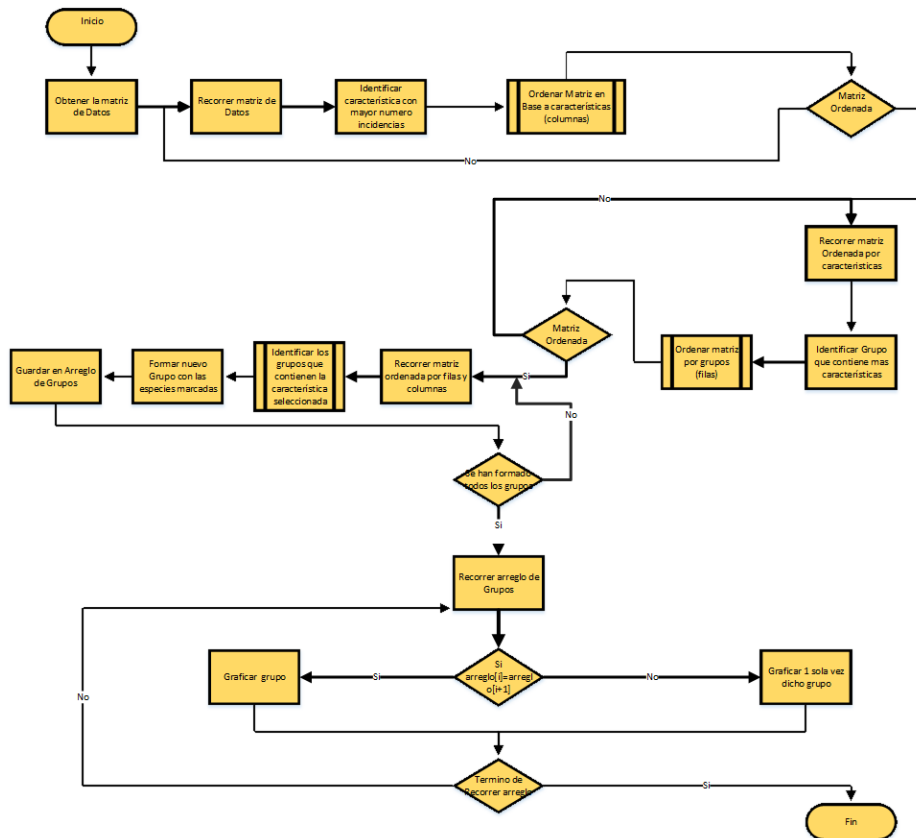


Figura 23 Diagrama de Flujo Hennig

5.2 Seudocódigo

1. Obtener matriz de Datos
2. Recorrer matriz de datos
3. Buscar característica con mayor incidencia
4. Copiar columna en nueva matriz
5. Terminar de recorrer matriz de datos

6. Recorrer matriz de columnas ordenadas
 7. Buscar grupo (fila) con mayor número de características
 8. Copiar fila a nueva matriz
9. Terminar de recorrer matriz ordenada por columnas
10. Recorrer matriz ordenada por columnas y filas
 11. Identificar que grupos se unen con cada característica
 12. Guardar grupos en Array
13. Terminar de recorrer matriz ordenada por columnas y filas
14. Recorrer array de grupos
 15. Si $array[i] = array[i+1]$
 16. Graficar solo una vez dicho grupo
 17. Sino
 18. Graficar ambos grupos
19. Terminar de recorrer array
20. Fin de ejecución

5.3 Complejidad Matemática de Hennig

Como se observa en la figura 23, el algoritmo programado requiere tres ciclos. Cada ciclo contiene anidados dos ciclos para poder recorrer la matriz $(A(n^3) + B(n^3) + C(n^3))$, por tanto, la complejidad es $O(n^3)$.

5.4 Metodología de programación Hennig

Diseñar Software a la medida no es una tarea fácil, de hecho, es una de las más arduas al momento de la creación y actualización. Para facilitar la reconstrucción un árbol filogenético es necesario que la información sea presentada de la mejor forma posible ya que esto ayuda directamente con el procesamiento de los datos.

Cuando se presenta una matriz de información dada directamente por el usuario, aun no presenta tratamiento alguno para comenzar a trabajar. Como paso inicial ordenamos el contenido de la matriz de tal forma que las filas se ordenen de mayor a menor, es decir, poner las filas que tienen más 1's al inicio de una nueva matriz de datos.

0	1	2	3	4	5	6	7	8	9	10
5	1	1	1	0	0	0	1	1	1	0
6	1	1	1	0	0	0	1	0	1	1
7	1	1	0	0	0	0	1	0	1	1
8	1	0	0	0	0	0	1	0	1	1
2	0	0	0	1	1	1	0	0	0	0
3	0	0	0	1	0	1	0	0	0	0
4	0	0	0	0	0	0	1	0	1	0
1	0	0	0	0	0	0	0	0	0	0

Figura 24 Matriz de Datos ordenada por filas

Basados en la matriz de la figura 24. Se identifica que columnas son las que tienen mayor número de 1's (el carácter que se presenta en el mayor número de especies) para después ordenarla tomando como base este principio (Fig. 25)

0	7	9	1	2	10	3	4	6	5	8
5	1	1	1	1	0	1	0	0	0	1
6	1	1	1	1	1	1	0	0	0	0
7	1	1	1	1	1	0	0	0	0	0
8	1	1	1	0	1	0	0	0	0	0
2	0	0	0	0	0	0	1	1	1	0
3	0	0	0	0	0	0	1	1	0	0
4	1	1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0

Figura 25 Matriz de datos ordenada por filas y columnas

Ahora bien, como se puede observar, se presenta una forma de ver que caracteres tienen más peso dentro de la reconstrucción. En la figura 25 se puede observar que el mayor número de 1's se concentra en el triángulo superior de nuestra matriz de datos.

Como resultado, esto ayuda a tener una idea más clara de que especies están en profundidad y cuales son aquellos que presentan menos profundidad en un árbol filogenético (Fig. 26).

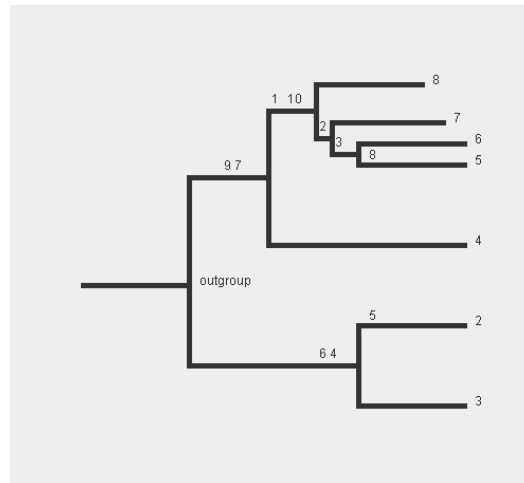


Figura 26 Cladograma generado con la aplicación desarrollada

6. Metodología Propuesta Simple LinkAge

El Análisis Clúster o Análisis de Conglomerados, es una técnica estadística multivariante que busca agrupar o separar elementos o variables tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos.

El Análisis Clúster tiene una tradición en muchas áreas de investigación. Sin embargo, las soluciones que se obtienen no son únicas, la medida de pertenencia al conglomerado para cualquier número de soluciones depende de muchos elementos que participan en el procedimiento elegido. Por otra parte, la solución clúster depende totalmente de las variables utilizadas, la adición o destrucción de variables relevantes puede tener un impacto substancial sobre la solución resultante. (Hernández, 2011) clasifica los conglomerados en dos categorías:

6.1 Algoritmos de partición

Método de dividir el conjunto de observaciones en k conglomerados (clusters), en donde k lo define inicialmente el usuario.

6.2 Algoritmos jerárquicos

Son métodos que entregan una jerarquía de divisiones del conjunto de elementos en n conglomerados, es decir, en base a un grupo de estudio puede unir o dividir dicho elemento en N conglomerados, las uniones o divisiones representan un orden jerárquico dentro del conglomerado final. Por esta razón los algoritmos jerárquicos a su vez pueden ser aglomerativos o disociativos (Fig. 27).

- Un método jerárquico disociativo sigue el sentido inverso, parte de un gran conglomerado y en pasos sucesivos se va dividiendo hasta que cada observación queda en un conglomerado distinto.
- Un método jerárquico aglomerativo parte con una situación en que cada observación forma un conglomerado y en sucesivos pasos se van uniendo, hasta que finalmente todas las situaciones están en un único conglomerado.

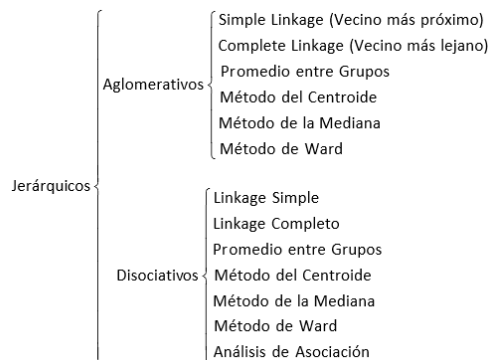


Figura 27 Clasificación de algoritmos jerárquicos

Para poder unir variables o individuos es necesario tener algunas medidas numéricas que caractericen las relaciones entre las variables o los individuos. Cada medida refleja una asociación en un sentido particular y es necesario elegir una medida apropiada dependiendo del problema que se esté tratando. Las medidas de asociación pueden ser de distancia o de similitud.

- Cuando se elige una distancia como medida de asociación, los grupos formados contendrán individuos parecidos de forma que la distancia entre ellos ha de ser pequeña.
- Cuando se elige una medida de similaridad, los grupos formados contendrán individuos con una similaridad alta entre ellos.

6.3 Etapas del Análisis de Conglomerados

Como cualquier algoritmo, es conveniente identificar los pasos que se requieren para efectuar el análisis. Los pasos dentro del análisis de conglomerados son:

1. Elección de variables
2. Elección de las medidas de asociación
3. Elección de la técnica de clúster

Dentro del análisis filogenético que a continuación se presenta, la elección de variables ha sido determinada por un experto que avala la veracidad de nuestra matriz de datos. Nuestra medida de asociación será calculada en base a la diferencia de valores que tendrán cada una de las características homólogas con respecto al conjunto de taxones que forman la matriz de datos. Estas diferencias serán calculadas y representadas en una matriz de distancias.

Como nuestro objetivo es formar un conglomerado único a partir de un conjunto de elementos, en cada observación (iteración) se ha de formar un conglomerado. Se repite ésta acción hasta que finalmente todos los taxones estén conectados en un único árbol. Por esta razón se escogió el algoritmo jerárquico aglomerativo Simple LinkAge (vecino más próximo) que nos ayudará a determinar la distancia respectiva entre especies con las demás pertenecientes al conjunto de datos a observar. Este método nos permite crear el árbol de una manera sencilla de tal forma que se necesite crear una matriz de distancias en la cual se marcan las distancias entre especies respectivamente. Una vez calculada dicha matriz, se identifica en qué lugar existe menos distancia para después unir las especies que tienen menor distancia entre sí.

Las distancias entre conglomerados son funciones entre las observaciones, y por ende hay varias formas de definir las. Usaremos la distancia mínima, también conocida como distancia con el vecino más cercano (Fig 28).

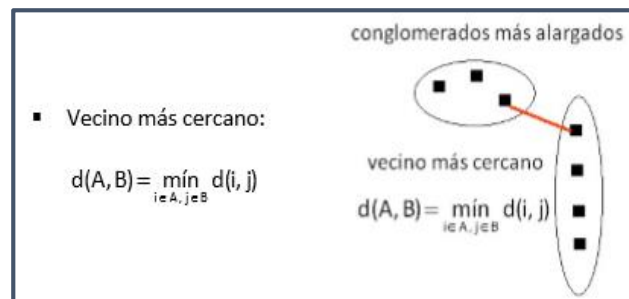


Figura 28 Distancia del vecino más cercano

Algoritmo:

- Se comienza con una matriz con n taxones (matriz de datos) y con una matriz $n \times n$ de distancias $\Delta = (\delta_{ij})$ simétrica y con ceros en la diagonal.
- Se busca en la matriz de disimilaridades los grupos que tengan menor distancia entre si (el par de grupos más próximos). Sean U y V los grupos más próximos, y $d(UV)$ su distancia.
- Se unen los grupos U y V , y se etiqueta el nuevo grupo como (UV) . Se actualiza la matriz de disimilaridades, de la siguiente forma:
 - a) Se borran las filas y columnas correspondientes a los grupos U y V .
 - b) Se añade una fila y una columna con las distancias entre el grupo (UV) y los grupos restantes.
- Repetir los pasos 2 y 3, $n - 1$ veces. Al final, todas las unidades estarán incluidas en un único grupo y las etiquetas de los grupos que se han unido, así como las distancias con las que se unieron (Hernández, 2011). La figura 29 representa el diagrama de flujo del algoritmo de Conglomerados Simple LinkAge

Para la realización del segundo paso, es necesario la definición de una medida de disimilaridad entre grupos. La medida de disimilaridad que se defina determina el tipo de método aglomerativo. Las disimilaridades que usaremos es la de simple Linkage o vecino más próximo, en este método, la disimilaridad entre dos grupos es la disimilaridad entre sus miembros más próximos, es decir, si U y V son dos grupos, entonces se define de la siguiente forma (Ecuación 6):

$$d_{uv} = \min\{d_{ij} \mid i \in U, j \in V\} \quad (6)$$

6.4 Diagrama de Flujo

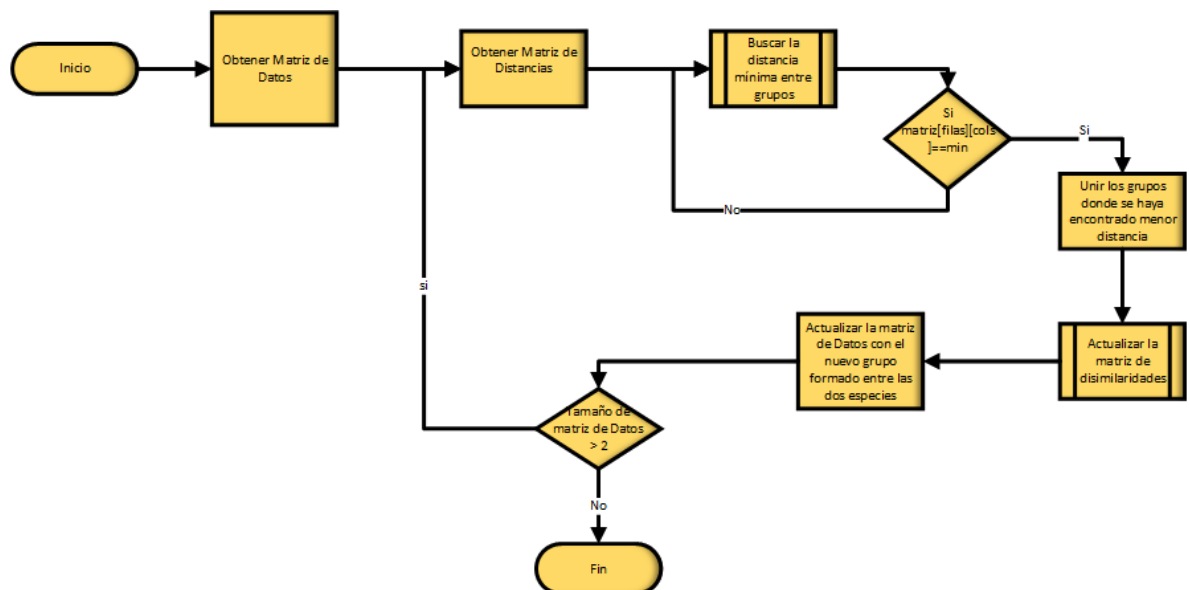


Figura 29 Diagrama de Flujo Conglomerados

6.5 Complejidad Matemática Simple Linkage

Como se observa en la figura 29, se tienen dos ciclos anidados para buscar la distancia mínima entre grupos y actualizar la matriz de disimilaridades ($A(n^2) + B(n^2)$), por lo que su complejidad es $O(n^2)$. Estos dos ciclos están inmersos en un ciclo. Por lo que el algoritmo tiene una complejidad $O(n^3)$.

6.6 Seudocódigo

1. Leer matriz de datos
2. Do
3. Calcular matriz de distancias
4. Recorrer matriz de distancias [filas][columnas]
5. Min=Buscar el valor mínimo ()
6. if matriz de distancias [filas][columnas] es igual min
7. Grupo1 = fila
8. Grupo2 =columna
9. Termina if
10. Termina ciclo
11. Unir grupo1 y grupo2
12. Actualizar la matriz de datos
13. While matriz de datos > 2

Para entender de mejor forma el funcionamiento de este algoritmo se presenta el siguiente ejemplo: Teniendo nuestra matriz de Datos (ver Tabla 7).

Tabla 7 Matriz de Datos (Autoría Propia)

	Características		
Taxón	1	2	3
A	1	1	1
B	0	1	1
C	0	1	0

Se calcula a distancia entre cada una de las especies. Para los cálculos habrá que observar el número de cambios que se tiene entre los taxones (ver Tabla 8)

Tabla 8 Matriz para calcular las diferencias entre especies (Autoría Propia)

A	1	1	1
B	0	1	1

La distancia entre A y B es únicamente de 1 ya que la primera característica es 1 en el grupo **A** y 0 en el grupo **B**. La característica 2 y 3 tienen el mismo valor por lo tanto no incrementa la distancia.

Sucesivamente, se realizan los cálculos entre todas las especies con el fin de obtener la matriz de distancias (ver Tabla 9).

Tabla 9 Matriz de Distancias (Autoría Propia)

Matriz de Distancias			
Distancia	A	B	C
A	0		
B	1	0	
C	2	1	0

Se observa que la matriz de distancias es una Matriz Simétrica puesto que $d(A, B) = d(B, A)$. Por lo tanto, no necesitamos completar la matriz ya que el triángulo inferior de la matriz contendrá los mismos valores que el triángulo superior y la diagonal principal de nuestra matriz de distancias siempre estará llena de ceros ya que $d(A, A) = 0$.

La distancia mínima la podemos ubicar entre el taxón B y el taxón C, por lo tanto, B y C forman un nuevo grupo (ver figura 30).

$$d(B, C) = 1$$

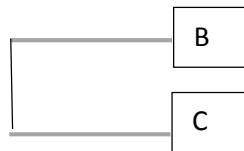


Figura 30 Conglomerado (Autoría Propia)

Ahora los grupos seleccionados dentro de la matriz de datos serán modificados como se ve en la matriz de datos (ver Tabla 10)

Tabla 10 Matriz de Datos (Autoría Propia)

Matriz de Distancias			
Taxón	Características		
	1	2	3
A	1	1	1
B-C	0	1	0

La nueva matriz de distancias está definida de la siguiente forma: (ver Tabla 11):

Tabla 11 Matriz de Distancias (Autoría Propia)

Matriz de Distancias		
Distancia	A	B-C
A	0	
B-C	2	0

$$D(A, B-C) = 2$$

Como se puede observar, se recalculan las matrices de distancias en base a la matriz de Datos y se une el grupo con su vecino más cercano (ver Figura 31).

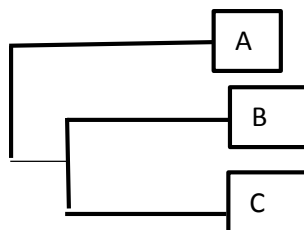


Figura 31 Conglomerado entre A-B-C (Autoría Propia)

7 Resultados y Funcionamiento del Sistema

7.1 Creación de un árbol filogenético en base al algoritmo de conglomerados Simple Linkage

La Tabla 12 es una matriz de Datos tomada de (Nápoles, 2015) la cual se usa como base para la creación de un cladograma usando el algoritmo Simple LinkAge

Tabla 12 matriz de Datos (Nápoles, 2015)

Ejemplo 3	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1 Pachymerus	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
2 Kythorinus	1	1	1	1	0	1	0	0	1	1	0	0	1	1	2	1	1	0	0	1
3 Megacerus	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	0	1	0	1	2
4 Callosobruchus	1	1	1	1	1	1	0	1	0	1	0	1	0	2	1	0	1	1	2	1
5 Conicobruchus	1	1	1	1	1	1	1	0	0	1	0	0	0	1	1	1	0	1	2	1
6 Rhipibruchus	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	1	1
7 Pectinibruchus	1	1	1	1	1	1	0	1	1	1	0	0	0	1	1	1	1	1	1	1
8 atrolineatus	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	2	1
9 walker	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	2	1
10 lumae	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	2	1

A partir de la tabla 12 se procede a calcular la primera matriz de Distancias. Debemos identificar dentro de nuestra matriz cuales son los taxones que tienen menor distancia entre sí (menores diferencias). Un dato importante a considerar dentro de la matriz de distancias es que pueden existir diferentes intersecciones entre especies que contienen la distancia

mínima 0 (cero). Es por este motivo que la cantidad de árboles incrementa en base al número de taxones que participan en el análisis como se especificó anteriormente (ver tabla 12).

Matriz de Distancias										
	1	2	3	4	5	6	7	8	9	10
1	0									
2	15	0								
3	11	6	0							
4	13	8	4	0						
5	12	7	4	5	0					
6	13	8	4	0	5	0				
7	14	5	5	3	4	3	0			
8	13	8	4	0	5	0	3	0		
9	13	8	4	0	5	0	3	0	0	
10	13	8	4	0	5	0	3	0	0	0

Tabla 13 Matriz de Distancias (Autoría Propia)

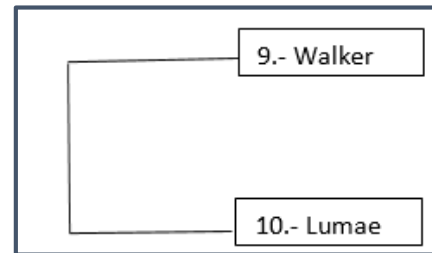


Figura 32 Conglomerado (Autoría Propia)

En la Tabla 13 se puede observar que los grupos que presentan menos distancia entre si son el grupo 910. Por tanto, estos dos grupos han formado un nuevo conglomerado que consta de la unión de los dos grupos mencionados, en la figura 32 se describe gráficamente esa unión.

2.- Recalculamos la matriz de Distancias (ver Tabla 14)

Matriz de Distancias										
	1	2	3	4	5	6	7	8	910	
1	0									
2	15	0								
3	11	6	0							
4	13	8	4	0						
5	12	7	4	5	0					
6	13	8	4	0	5	0				
7	14	5	5	3	4	3	0			
8	13	8	4	0	5	0	3	0		
910	13	8	4	0	5	0	3	0	0	0

Tabla 14 Matriz de Distancias (Autoría Propia)

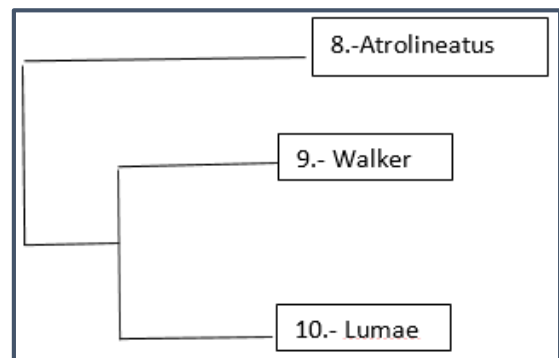


Figura 33 Conglomerado (Autoría Propia)

El siguiente elemento a unir está dado por el grupo 8. En la tabla 14 se puede dar cuenta que efectivamente los grupos que comparten menos distancia entre si son el grupo 8 y el grupo 910. Por tanto, deberán unirse para formar el grupo 8910 (Fig. 33).

3.- Se recalcula la matriz de distancias (Ver tabla 15):

Matriz de Distancias								
	1	2	3	4	5	6	7	8910
1	0							
2	15	0						
3	11	6	0					
4	13	8	4	0				
5	12	7	4	5	0			
6	13	8	4	0	5	0		
7	14	5	5	3	4	3	0	
8910	13	8	4	0	5	0	3	0

Tabla 15 Matriz de Distancias (Autoría Propia)

Ahora bien, como se ha visto, las matrices de distancias deben ser calculadas cada que se crea un nuevo grupo. En la tabla 15 se puede observar que el grupo 6 e intersección con el grupo 8910 son los que tienen la mínima distancia entre sí. Existen otras intersecciones que con diferentes grupos los cuales podrían formar nuevos conglomerados. Sin embargo, en nuestro ejemplo se tomará el valor de intersección mínima que se encuentra más en profundidad.

4.-Se recalcula la matriz de Distancias obteniéndose la tabla 16:

Matriz de Distancias							
	1	2	3	4	5	7	68910
1	0						
2	15	0					
3	11	6	0				
4	13	8	4	0			
5	12	7	4	5	0		
7	14	5	5	3	4	0	
68910	13	8	4	0	5	3	0

Tabla 16 Matriz de Distancias (Autoría Propia)

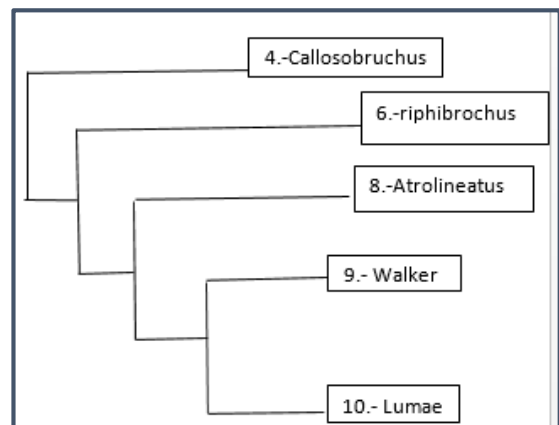


Figura 34 Conglomerado (Autoría Propia)

El siguiente grupo a formar estará dado por el grupo 4 y el grupo 68910, de esta forma nuestro nuevo grupo se llamará 468910 que son los grupos que se han ido formando. Dicha representación se ve reflejada en el conglomerado 4 (Fig. 34).

5.- Recalculando las distancias se agrega el grupo 7 al grupo 468910 debido a que en la matriz de distancias (ver tabla 17) puede verse que son los grupos en los cuales existe menor distancia. Su representación se puede ver en la Figura 35.

Matriz de Distancias						
	1	2	3	5	7	468910
1	0					
2	15	0				
3	11	6	0			
5	12	7	4	0		
7	14	5	5	4	0	
468910	13	8	4	5	3	0

Tabla 17 Matriz de Distancias (Autoría Propia)

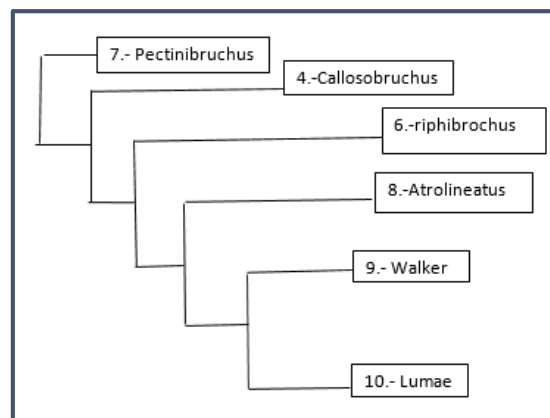


Figura 35 Conglomerado (autoría propia)

6.- Se procede a hacer los cálculos para la siguiente matriz de Distancias obteniendo como resultado la tabla 18.

Matriz de Distancias					
	1	2	3	5	7468910
1	0				
2	15	0			
3	11	6	0		
5	12	7	4	0	
7468910	11	7	3	4	0

Tabla 18 Matriz de Distancias (Autoría Propia)

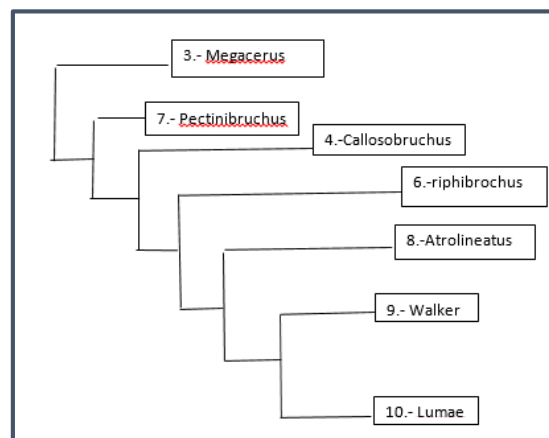


Figura 36 Conglomerado (Autoría Propia)

El nuevo conglomerado se forma uniendo los siguientes grupos que tengan la menor distancia entre sí. En la matriz de Distancias calculadas (ver tabla 18) como en todas las matrices de distancias anteriores, se debió calcular la matriz de distancias en base a la matriz de datos. En este punto la matriz deja observar que el grupo 3 debe unirse al conglomerado ya formado, su representación puede ser vista en el conglomerado (Fig. 36).

7.- Recalculamos la matriz de Distancias

Matriz de Distancias				
	1	2	5	37468910
1	0			
2	15	0		
5	12	7	0	
37468910	10	5	4	0

Tabla 19 Matriz de Distancias (Autoría Propia)

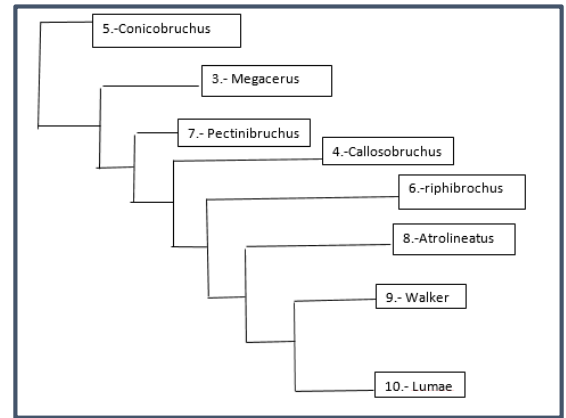


Figura 37 Conglomerado (autoría propia)

Se repite el proceso para cada iteración agregando los grupos con menor distancia entre sí. En esta iteración el grupo 5 se une al conglomerado como se puede observar en la figura 37.

8.- Para concluir, se calcula la matriz de Distancias final para terminar de unir nuestros taxones y formar conglomerado general

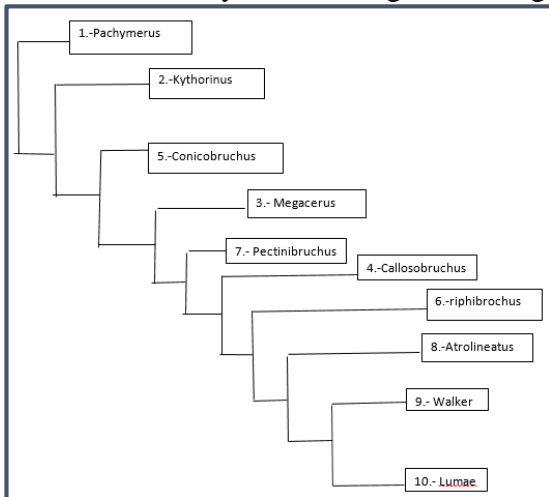


Figura 38 Conglomerado (Autoría Propia)

Matriz de Distancias			
	1	2	537468910
1	0		
2	15	0	
537468910	9	6	0

Tabla 20 Matriz de Distancias (Autoría Propia)

Nuestro conglomerado hasta este punto está casi finalizado. En este momento ya tenemos el grupo 537468910 al cual se une el grupo 2 (ver Figura 38)

A medida que la matriz de distancias decremento y el grafo incrementa, es decir, cuando se crea un nuevo conglomerado, se tiene la posibilidad de identificar cuales grupos son los que tienen menor diferencias y unirlos entre sí.

Cuando un taxón en comparación con otro taxón (ambos pertenecientes al grupo de estudio) encuentra menor distancia se dice que ha encontrado a su vecino más próximo.

En la figura 39a y 39b se muestran dos cladogramas finales basados en la misma matriz de datos. Se puede observar la similitud que presentan los dos cladogramas entre sí. Es decir, la matriz de datos sigue siendo la base para el proceso de creación del cladograma, sin embargo, los métodos de reconstrucción de estos son totalmente diferentes.

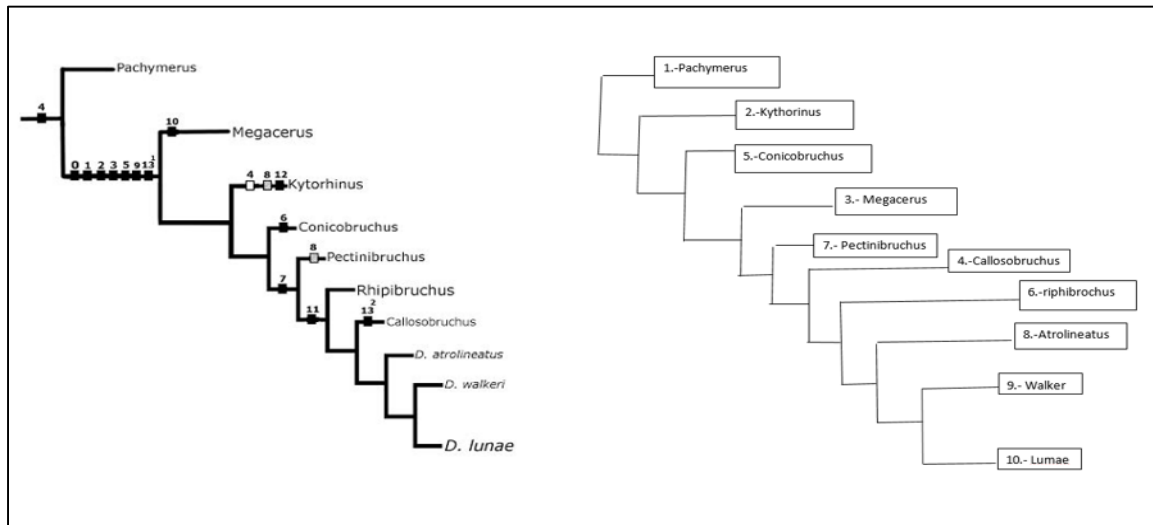


Figura 39 Árbol filogenético a) creado con el algoritmo de Hennig (Nápoles, 2015) b) creado con el algoritmo de conglomerados

7.2 Funcionamiento del sistema

La aplicación fue desarrollada en base al lenguaje de programación Java el cual, es independiente de plataforma. A continuación, se presentan las pantallas funcionamiento del sistema.

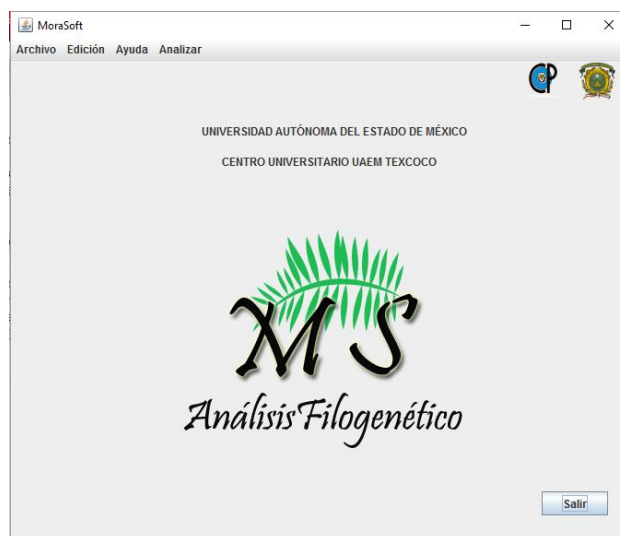


Figura 40 Pantalla inicial de la aplicación

Para la creación de un nuevo proyecto podemos ver la facilidad para elegir el número de características y especies que participaran en el análisis (Fig. 41).

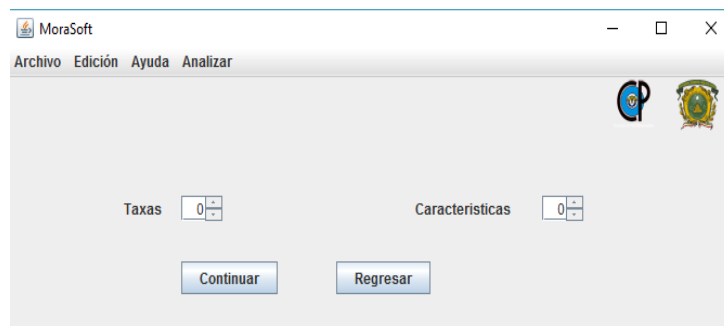


Figura 41 Seleccionar número de Especies y Características

Ahora, una vez seleccionadas las características y especies procedemos a crear una tabla para la inserción de datos (Fig. 42)

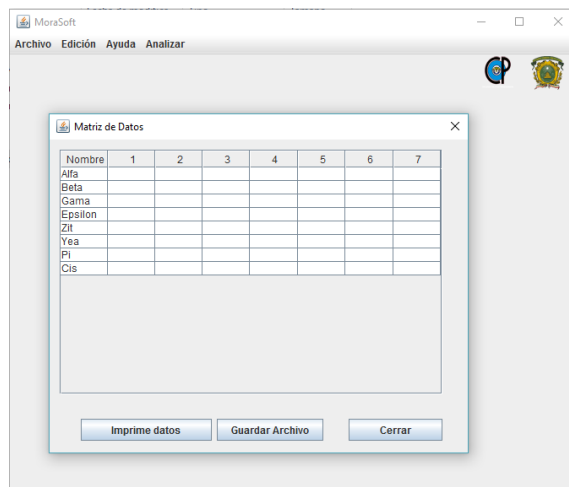


Figura 42 Interfaz para la captura de información

Una vez creada nuestra tabla de datos, seleccionamos el archivo y cargamos los datos que deben ser analizados (Fig. 43).

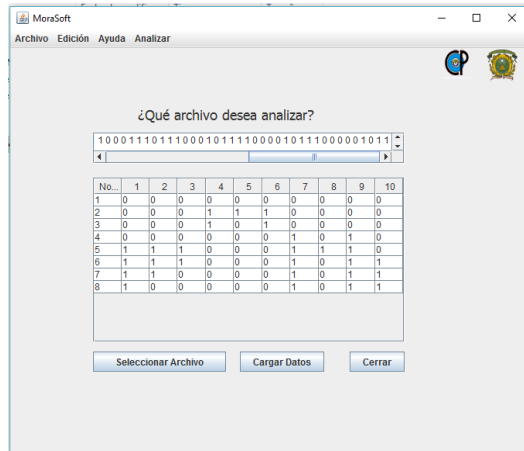


Figura 43 Tabla de Datos a ser analizados

El siguiente panel nos muestra como los datos han sido cargados y analizados (Fig. 44).

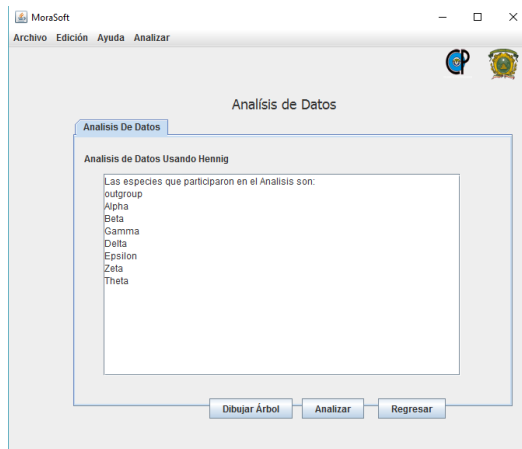


Figura 44 Datos analizados

Presentación del resultado final (Fig. 45)

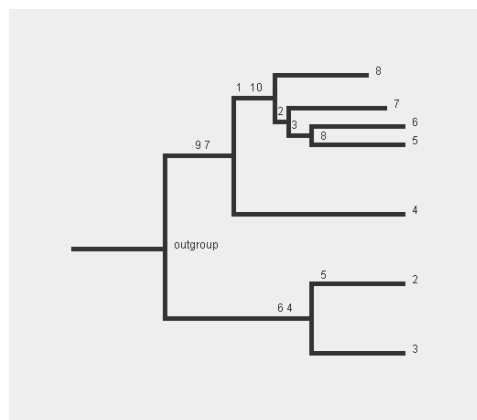


Figura 45 Resultado final

Discusión

El desarrollo de nuevas tecnologías y el incremento de las relaciones de las ciencias naturales con el cómputo han abierto la brecha para el desarrollo de software a la medida. Los lenguajes de programación son utilizados como el medio por el cual se pueden plasmar ideas, algoritmos y llegar a un fin común deseado ya que un mismo sistema puede ser implementado de forma local o en un servidor web para uso público.

Por ende, la entomología y el cómputo tienen una importancia suprema dentro de muchas áreas de investigación lo cual no separa al conocimiento sino antes bien logra la fusión de dos ciencias aparentemente excluyentes entre sí.

La importancia de la representación de los datos finales se vuelve indispensable al momento de comenzar a programar. Los algoritmos propuestos tienen diferencias muy marcadas entre sí; estas diferencias pueden ser vistas en mayor amplitud en base a que el algoritmo de Hennig da la oportunidad de identificar qué carácter o conjunto de caracteres se utilizan para crear una unión entre especies mientras que Simple linkage une en cada iteración un grupo a la vez sin representar los caracteres que han sido tomados para unir cada grupo y esto se debe a que no es un algoritmo basado en caracteres sino en distancias.

Por este motivo la programación de Hennig tiene muchos criterios a considerar. La estructura que se toma como base es a partir de una matriz de datos que el usuario introdujo, con dicha matriz buscamos invertir el orden de las filas tomando en cuenta que los datos originales son números. El siguiente paso es ordenar nuevamente la matriz por columnas y como paso final dentro de un arreglo plasmar los grupos que tienen el o los caracteres para poder unir los grupos. Si existieran dos elementos del arreglo con los mismos valores se concatenan las especies que unen a los grupos para tener las relaciones existentes y así evitar que se duplique información.

Cada elemento del arreglo se ha representado dentro de 3 matrices de coordenadas que se utilizará para su posterior traficación, la primera matriz estará ocupada por aquellos grupos que solo tienen una característica o dicho de otra forma por los grupos en los cuales la característica x es la única que se hace presente. La segunda matriz de coordenadas tiene como base aquellas uniones donde sólo participan dos especies para cada característica, es decir, buscamos los pares que se pueden hacer en base a las características seleccionadas y finalmente la matriz 3 es aquella matriz en la cual están los grupos que contienen 3 o más especies por cada carácter.

Al utilizar este tipo de matrices de coordenadas se puede hacer una búsqueda en de elementos ya creados. Para tener un mejor control dentro de las matrices de coordenadas, en específico en la matriz de coordenadas 3, es importante que sea ordenada de forma ascendente de acuerdo a la longitud de los grupos que están dentro de ella. Ya teniendo ordenada la matriz de coordenadas para cada especie se hace el mismo procedimiento de buscar dentro de las matrices de coordenadas 1 y 2 la existencia de grupos que estén contenidos en este nuevo grupo a formar.

De esta manera se busca la graficación de todos los elementos que participan dentro de la matriz de coordenadas, en párrafos anteriores se dijo que la matriz de datos

originalmente trabaja con números enteros, pero al momento de la creación de las matrices de coordenadas se trabaja con matrices del tipo String y esto debido a la siguiente razón; si tuviésemos un grupo formado por alguna característica que tuviese como longitud de tamaño 10, el compilador marca un error de desbordamiento ya que el tipo de datos entero no soporta números con esta longitud, por esta razón se decidió guardar la información de las matrices como tipo Sting, de esta forma puede crecer en tamaño sin importar la cantidad de elementos que tenga un grupo. El problema derivado de esto es que en cada operación se realizan constantes conversiones entre los tipos de dato Sting e int.

Por lo que se determina que la dificultad para programar el algoritmo de Hennig es alta debido a todos los arreglos de datos y la alta recursividad que se maneja para encontrar las relaciones para cada grupo perteneciente a la matriz de coordenadas.

Conclusiones

El análisis filogenético dentro de la entomología es un proceso que requiere la validación y el trabajo por el experto. Las relaciones entre los taxones deben estar basadas en las similitudes que deben tener un árbol filogenético que nos permite reconstruir las similitudes entre los organismos.

La construcción de los árboles filogenéticos o cladogramas se puede llevar a cabo por diferentes metodologías que se basan en principios matemáticos probabilísticos y que determinan cada una de las diferentes formas de la reconstrucción. Por esta razón, el uso de grupos dentro del campo entomológico facilita al experto el análisis filogenético, porque se puede utilizar como una herramienta de base para generar relaciones genealógicas entre las especies. Es conveniente decir que los resultados pueden ser muy similares entre sí, porque se basan en la matriz de datos original que trata de vincular las características más esenciales para su análisis.

El análisis de conglomerados utiliza muchos algoritmos que se pueden utilizar conjuntamente para ofrecer resultados diferentes, por lo que su uso dentro de la economía, estadística y muchas disciplinas es esencial para la representación de datos estructurados.

No hay estandarización en la formulación de cladogramas ya que depende de los métodos utilizados como se mencionó anteriormente. Las matrices de datos pueden tener diferentes características a causa de los grupos de estudio que han de ser analizados.

En la parte de computacional, aunque la complejidad algorítmica en ambos algoritmos tiene un orden $\mathcal{O}(n^3)$, se muestra que el algoritmo Simple Linkage reduce la complejidad para su comprensión y la programación. Además de que el algoritmo propuesto por Hennig es mucho más complejo en la graficación.


Bibliografía


- Bastar, S. G. (2012). *Metodología de la Investigación*. Estado de México : Red Tercer Milenio .
- Booch, G., & Rumbaugh, J. (s.f.). *El lenguaje Unificado de Modelado*. Recuperado el 7 de Julio de 2016, de [elvex.ugr.es: http://elvex.ugr.es/decsai/java/pdf/3E-UML.pdf](http://elvex.ugr.es/decsai/java/pdf/3E-UML.pdf)
- Goyenechea, I. (2006). Sistemática: su historia, sus métodos y sus aplicaciones en las serpientes. *Instituto de Ciencias Básicas e Ingeniería, Área Académica de Biología, Universidad Autónoma del Estado de Hidalgo.*, 9. Recuperado el 20 de 11 de 2014, de <http://entomologia.rediris.es/documentos/taxonomia.htm>
- Hernández, S. d. (2011). *Análisis de Conglomerados*. Madrid, España: Universidad Autonoma de Madrid.
- Lara, S. L. (2015). El método que nos une: el empleo de la cladística en Antropología . *Desde el Herbario CICY* .
- Linnaeus, C. (1735). *Systema Naturae*. Sweden.
- Lipscomb, D. (1998). *Basics of Cladistic Analysis*. Washington D. C.: George Washington University.
- Lopez Razo, B. S., Ayala de la Vega, J., Lugo Espinoza, O., & Napoles Romero , J. (2016). Cluster Analisys as a methodology within Phylogenetic Systematics to Construct Phylogenetic Trees. *International Journal of Modern Engineering Research*, 15.
- Morrone, J. J. (2000). *EL lenguaje de la Cladística*. Mexico: UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO.
- Napoles, D. J. (1990). *Entomologia Sistemática*. ColPos.
- Nápoles, J. R. (2015). Systematics of the seed beetle genus *Decellebruchus* Borowiec, 1987 (Coleoptera: Bruchidae).
- Ramos, A. C. (2007). *La sistemática, base del conocimiento de la biodiversidad*. Pachuca, Centro, Hidalgo: Universidad Autónoma del Estado de Hidalgo .
- Rodriguez Baena, L. (s.f.). *www.colimbo.net*. Recuperado el 03 de Julio de 2016, de http://www.colimbo.net/documentos/documentacion/fipo/Maquetar_con_CSS.pdf
- Rodriguez Catalán, P. (Septiembre de 2001). *Anàlisis Filogenètics*. Recuperado el 05 de Septiembre de 2015, de [academia.edu: http://www.academia.edu/3578130/AN%C3%81LISIS_FILOGEN%C3%89TICOS](http://www.academia.edu/3578130/AN%C3%81LISIS_FILOGEN%C3%89TICOS)
- Tato Gomez , A. (2011). *Grupo de Bioinformática de la Facultad de Matemáticas*. Recuperado el 04 de Octubre de 2015, de <http://mathgene.usc.es/cursoverano/cv2005/materiales/filogenia/filogenia1.pdf>
- Terrádez Gurrea, M. (s.f.). <http://www.uoc.edu/in3/emath/docs/Cluster.pdf>. Recuperado el 07 de Julio de 2016, de <http://www.uoc.edu/in3/emath/docs/Cluster.pdf>.


Washington, U. o. (s.f.). *Phylogeny Programs*. (University of Washington) Recuperado el 14 de Julio de 2015, de <http://evolution.genetics.washington.edu/phylip/software.html>

Anexos

Search by Title or ISSN:

Select language 

INDEX COPERNICUS
INTERNATIONAL 



[Home](#) ⇒ [Journal passport](#) ⇒ [Journal content](#)

International Journal of Modern Engineering Research (IJMER) ISSN: 2249-6645

[ICI Journals Master List 2014](#)
Now available! Annual Report ICI Journals Master List 2014 summarizing the 2014 year with full list of journals and publishers from database of Index Copernicus.

[Index Copernicus Search Articles](#)

Volume 5, 2015



Cluster Analysis as a Methodology Within Phylogenetic Systematics to Construct Phylogenetic Trees

Benito Samuel López Razo¹, Joel Ayala De La Vega², Oziel Lugo Espinosa³, Jesús Romero Nápoles⁴

^{1, 2, 3}(C. U. UAEM Texcoco, Universidad Autónoma del Estado de México, México)

⁴(Departamento de Entomología, Colegio de Posgraduados, México)

I. INTRODUCTION

All evolutionary studies of groups of species are based on the choice of appropriate characteristics for rebuilding their phylogenies (a phylogeny is the relationship or kinship among species in general and tries to reconstruct evolutionary relationships). A phylogenetic analysis reconstructs the evolutionary relationships between species, which descend from common ancestors and, furthermore, which are the genetic distances or separation times between these species [1].

To generate a phylogenetic analysis characters must have two requirements: independent of each other and be homologous, they have the same origin and the same function in all organisms Study

The nature of those characters can be varied. Any source of validated and proved phylogenetic information can provide characters for an evolutionary study. Among the main evolutionary studies that have been developed stand two methods: The methods that have been taken as morphological characters base in which the presence of physical characteristics that describe the species is identified, and methods that have been based on molecular characteristics as the sequence DNA [2]. These characters are recorded in a data matrix within which, the state in which the character has been observed is represented with zero if it is absent or one if present respectively, and whether it is a character that may be present in the species with different values (multi-state) within the data matrix can be represented by the value corresponding to that character [2].

For this reason, homologous characters, once they have been validated and proven, may be taken as the basis for an evolutionary study because they provide enough information for the reconstruction of a phylogenetic tree.

II. DATA MATRIX

The species to be analyzed are defined based on the each scientist interested group. Therefore, once the data set has been obtained, it needs to be translated into a structure that allows fully represent the relationships with each other.

For this, it is common to find the data represented by a matrix in which taxa (species) are grouped in rows and the characters in columns.

An important component to know is the term external group (out group) whose main function is that it can be used as a comparison group on which we could take it as a base to do the math measurement and comparison to join groups each other and to entrench the resulting cladogram. If we do not include the out group within our data matrix the cladogram will lack the root. The matrix (1) shows an example of a data set containing the group out.

Matrix 1. Example of a data matrix (Own creation)

Matriz de Datos				
	1	2	3	4
Out Group	0	0	0	0
Taxon 1	0	0	1	0
Taxon 2	1	0	0	1
Taxon 3	1	1	0	0
Taxon 4	1	1	0	1

Matrix (2) presents real data, in which a set of 10 taxa is shown and each taxon has 21 features.

Matrix 2. Real Data Matrix [3]

1	Psittacus	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Kittiwake	1	1	1	1	0	1	0	0	1	1	0	0	1	1	2	1	1	0	0	1
3	Megascops	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	0	1	0	1	2
4	Callipepla	1	1	1	1	1	1	0	1	0	1	0	1	0	2	1	0	1	1	2	1
5	Centurus	1	1	1	1	1	1	0	0	1	0	0	0	1	1	1	0	1	2	1	1
6	Rhipidura	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	1	1
7	Psittacus	1	1	1	1	1	1	0	1	1	1	0	0	0	1	1	1	1	1	1	1
8	Atalapha	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	2	1
9	Walker	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	2	1
10	Junco	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	2	1

III. CLADOGRAMS

A cladogram is a diagram of data as a tree reflecting the genealogical relationships of terminal taxa [4]. Phylogenetic trees or cladograms could be rooted or not. The Rooted trees have a particular node called root from which begins to come off the evolutionary path that is formed. A tree not rooted specifies the relationships among taxa but does not define the evolutionary path see Fig. (1)

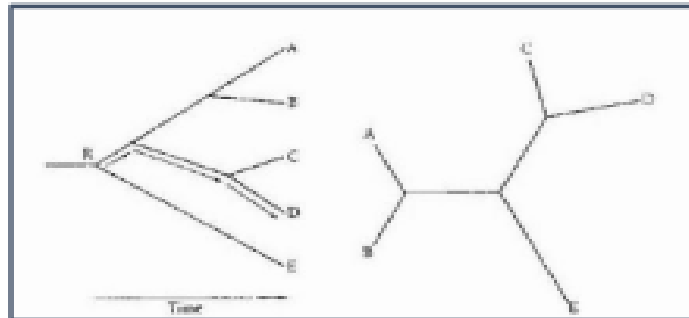
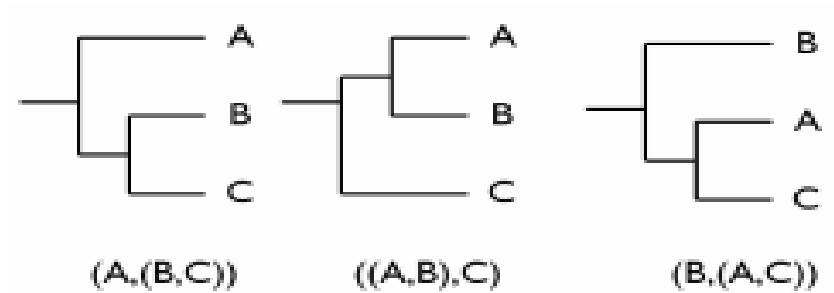


Figure 1 Root tree and not root tree [5]

Based on the number of taxa to be used in the study will be a wide variety of trees. For example, if we talk about 3 species A, B, C. there may be three rooted trees and one without roots see Fig(2).



Figures2. Tree with three taxon [6]

According to [2] the possible number of rooted trees for n taxa can be calculated from, (1):

$$N_r = \frac{(2n-3)!}{2^{n-2}(n-2)!} \quad \text{Para } n > 2$$

Equation 1 Calculate root trees

Where:

N_r is the number of rooted trees.
 n is the number of taxa.

And the number of unrooted trees can be calculated by, (2):

$$N_u = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

Equation 2 Calculate unrooted trees

Where:

N_u is the number of trees without roots.
 n is the number of taxa used.

The number of possible trees rooted with n taxa is equal to the number of trees unrooted for $n-1$ taxa, the number of trees increases as n increases. Thus, from 12 species becomes difficult to quantify the number of trees with and without root that could be obtained (because it is an intractable problem since the compute all possible trees has a very high computational cost). For example, a year has 31,536,000 seconds, a Pentium IV processor executes four million instructions per second, which runs about 126.144×10^9 instructions per year. Assuming a tree in each instruction is performed, and leaning in Figure 3, for 20 species it would take 65 011 380 years to show all the trees and for 30 species will take 3.925×10^{25} years [6].

Thus, when n is large, the expert can't analyze all the trees generated, as only one of those trees correctly represents the true evolutionary relationship. Therefore, heuristics that can generate the correct trees are used.

Especies	Número de árboles
1	1
2	1
3	3
4	15
5	105
6	945
7	10,395
8	135,135
9	2,027,025
10	34,459,425
11	654,729,075
12	13,749,310,575
13	316,234,143,225
14	7,905,853,580,625
15	213,458,046,676,875
16	6,190,283,353,629,375
17	191,898,783,962,510,625
18	6,332,659,870,762,850,625
19	221,643,095,476,699,771,875
20	8,200,794,532,637,891,559,375
30	4.9518×10^{26}
40	1.00905×10^{27}
50	2.75292×10^{28}

Figure 3 Number of trees depending the number of species [8]

IV. Methods for constructing phylogenetic trees.

Once we have determined the group of species and created the data matrix, we can start to construct phylogenetic trees through different methods, however all this, it is necessary to say that you can build many trees and each of these will constitute a different evolutionary hypothesis [7].

A part to consider into cladistics is that it is not an intuitive system, but is based on empirical methods of reconstructing using strict rules for example the common ancestors linked through synapomorphies. For this reason [5] classifies empirical methods used to reconstruct phylogenies follows:

Based on Distances

Algorithmic way:

- ultra-metric (UPGMA Unweighted Pair Group Method With Arithmetic mean)
- Additives (Neighbor Joining)
- Optimization form
- The relationship between neighbors (Neighborliness)
- Distances transformed

Based on characters:

- By optimization criteria:
- Hennig
- Maximum Likelihood (ML – Maximum Likelihood)
- Parsimony (Maximum Parsimony) Estimation of the goodness of reconstruction using analytical techniques and resampling (Bootstrap, Jackknife, Decay).

Phylogenetic systematics or cladistics was proposed by German entomologist Willi Hennig in 1950 to make phylogenies with a methodology that was testable and repeatable which, until that date could not be done. There was not a way to follow, instead of, the investigator's experience said which groups were more similar each other [8]. Therefore, the proposed method will be compared with Hennig algorithm, since today several entomologist still rely with such method.

V. HENNIG

Hennig's argument considers the information of each character one at a time. It is easy to understand by a small example taken from [9]. In this example we will take as a base a data matrix containing 4 study groups (taxa) and 6 characteristics (Matrix 3)

Matrix 3 Hennig Data Matrix [9]

	Characters					
	1	2	3	4	5	6
Out group	0	0	0	0	0	0
A	1	0	0	0	0	1
B	1	1	0	1	1	0
C	1	0	1	1	0	0

1. Character 1 unites the taxes (groups) A, B and C because they share the apomorphic characteristic 1 (Fig. 4).

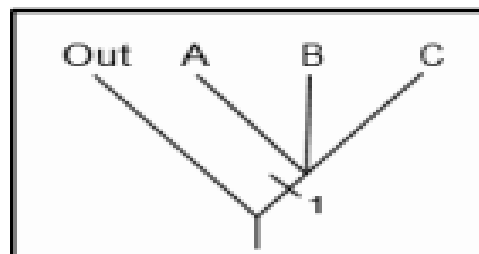


Figure 3. Tree with character 1 [9]

2. Character 2 - the derivative character is found only in the taxon B, and does not provide much information about the relationships between taxa (Fig. 5).

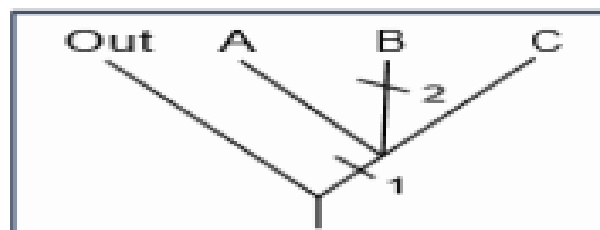


Figure 4 Tree with character 2 [9]

3. Character 3 - the derivative character is autapomorphic for group C (Fig. 6).

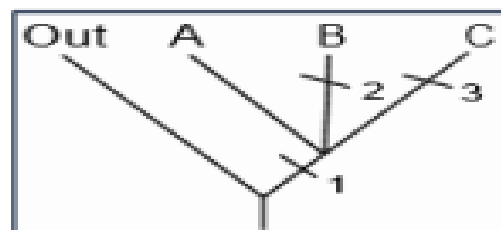


Figure 5 Tree with character 4 [9]

4. Character 4 - The derivative character is synapomorphic and unites the taxa A and B (Fig. 7)

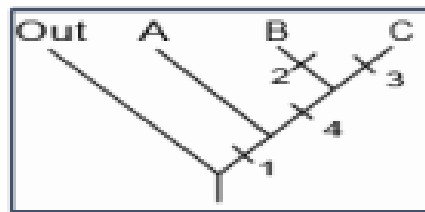


Figure 6 Tree with character 4 [9]

5. Character 5 - The derivative character is an autapomorphy for the taxon A (see Figure 8)

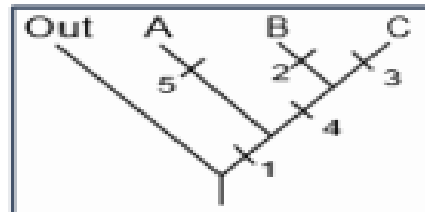


Figure 7 Tree with Character 5 [9]

The real data matrices are rarely so simple. However, the concept is the same. Within the following flow chart the function of our algorithm is shown (Fig. 9)

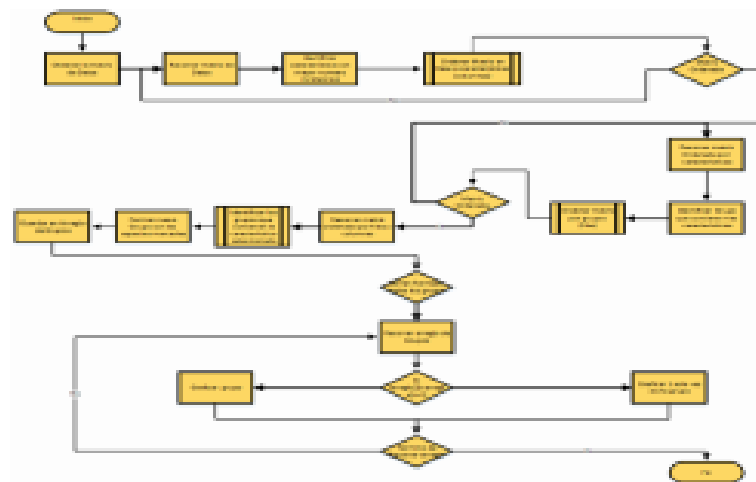


Figure 8 Hennig Flowchart

As seen in Figure 9, the programmed algorithm requires three cycles. Each cycle contains nested two cycles to tour the matrix ($A (n^3) + B (n^3) + C (n^3)$). Therefore the complexity is $O(n^3)$.

VI. CLUSTER

Cluster Analysis is a multivariate statistical technique that seeks to separate or link elements or variables trying to achieve maximum homogeneity in each group and the biggest difference between the groups.

Cluster analysis has a tradition in many areas of research. However, the solutions obtained are not unique, for cluster membership for any number of solutions depends on many elements involved in the procedure chosen. Moreover, the cluster solution depends entirely on the variables used, the addition or destruction of relevant variables can have a substantial impact on the resulting solution. [10], classifies into two categories the conglomerates.

Partition Algorithms

Method of dividing the set of observations in k clusters, where k initially is set by the user.

Hierarchical Algorithms

They are methods that provide a hierarchy of divisions of a set of elements in n clusters, i.e., based on a study group can unite or divide such element in N conglomerates, the unions or divisions represent a hierarchical order in the final conglomerate. Therefore, hierarchical algorithms in turn can be agglomerative or dissociative (Fig 10).

- A dissociative hierarchical method follows the reverse direction, part of a large conglomerate and is dividing successive steps until each observation is in a different cluster.
- An agglomerative hierarchical method starts with a situation where each observation forms a conglomerate and successive steps are joining, until finally all situations are in a single cluster.



Figure 9 Hierarchical classification algorithms [10]

To bind variables or individuals we needed to have some numerical measures that characterize the relationships between the variables or individuals. Each measure reflects a partnership in a particular sense and is necessary to choose an appropriate measure depending on the problem being treated. The measures of association could be distance or similarity.

- When you choose a distance as a measure of association, the groups will contain similar individuals formed so that the distance between them must be small.
- When a similarity measure is chosen, the groups formed contain individuals with high similarity between them.

Stages of Cluster Analysis

Like any algorithm, you should identify the steps required for the analysis. The steps within the cluster analysis are:

1. Select variables
2. Choose the measures of association
3. Election of the cluster technique

Within the phylogenetic analysis presented below, the choice of variables has been determined by an expert who supports the accuracy of our data matrix. Our measure of association will be calculated based on the difference of values that have each of the homologous characteristics with respect to the set of taxa that form the data matrix. These differences will be calculated and represented in a distance matrix. As our goal is to form a single cluster from a set of elements, in each observation (iteration) has to form a conglomerate. This action is repeated until finally all taxa are connected in a single tree. For this reason the agglomerative hierarchical algorithm Simple LinkAge was chosen (nearest neighbor) and will help us to determine the respective distance between other species belonging to the dataset was chosen to observe. This method allows us to create the tree in a simple manner. We need to

create a matrix of distances in which the distances between species are marked respectively. Once said matrix calculated, we identify where there is less distance to join the species whose distance is less.

The distances between clusters are functions between observations, and therefore there are several ways to define them. We will use the minimum distance, also known as distance to the nearest neighbor (Fig. 11).

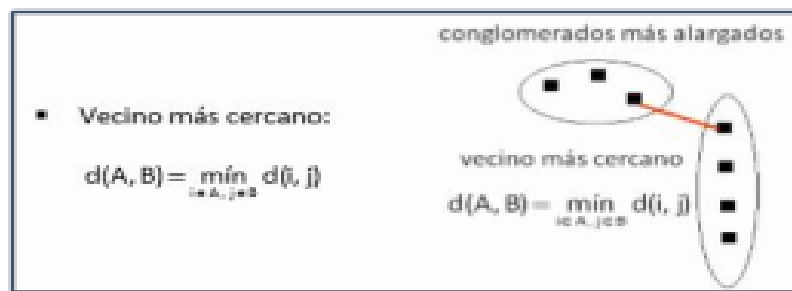


Figure 10. Neighbor nearest distance [10]

Algorithm:

- Starts with a matrix with n taxa (data matrix) and a matrix $n \times n$ distances $\Delta = (\delta_{ij})$ symmetrical with zeros on the diagonal.
- Groups with less distance between them (the two closest groups) is sought in the matrix of dissimilarities. Let U and V closest groups, and d (UV) its distance.
- U and V groups are joined, and the new group as (UV) is labeled. The dissimilarity matrix is updated as follows:
 - a) The rows and columns corresponding to the U groups and V are erased.
 - b) One row and one column with the distances between the group (UV) and the remaining groups is added.
- Repeat steps 2 and 3, $n - 1$ times. In the end, all units will be included in a single group and labels have joined groups and distances with which joined (Hernandez, 2011). 11 shows the flowchart of the algorithm Conglomerates Simple LinkAge

For carrying out the second step, requires the definition of a measure of dissimilarity between groups. The dissimilarity measure that is defined determines the type of agglomerative method. The dissimilarities we will use is *the single Linkage* or *nearest neighbor*, in this method, the dissimilarity between two groups is difference among its closest members, i.e., if U and V are two groups, then defined as follows, (3):

$$d_{uv} = \min\{d_{ij} \ 0: i \in U, \quad j \in V\}$$

Equation 3 Simple Linkage dissimilarity measure

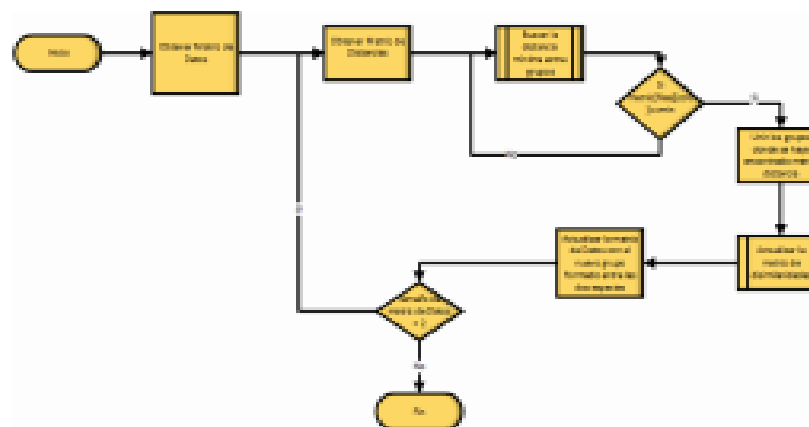


Figure 11 Conglomerate flowchart

As shown in Figure 12, there are two nested loops to find the minimum distance between groups and update the dissimilarities matrix ($A (n^2) + B (n^2)$), so that its complexity is $O (n^2)$. These two cycles are immersed in a cycle. So the algorithm has a $O (n^3)$ complexity. To better understand how this algorithm works, see the following example: With our Data Matrix (Matrix 4).

Matrix 4 Data Matrix (own creation)

		Characters		
Taxón		1	2	3
A		1	1	1
B		0	1	1
C		0	1	0

The distance between each of the species is estimated. For calculations will have to observe the number of changes that have among taxa (Matrix 5)

Matrix 5. Matrix to calculate the distances between 2 groups (own creation)

A	1	1	1
B	0	1	1

The distance between A and B is only 1 because the first characteristic is one in the group A and 0 in group B. The characteristic 2 and 3 have the same value therefore does not increase the distance.

Successively, the calculations between all species in order to obtain the distance matrix (Matrix 6) are performed.

Matrix 6 Distances Matrix (own creation)

Distance Matrix			
Distancia	A	B	C
A	0		
B	1	0	
C	2	1	0

As we can see the distance matrix is a Symmetric matrix since $d (A, B) = d (B, A)$. Therefore, we don't need complete the matrix because the lower triangle of the matrix contain

the same values as the upper triangle and diagonal of our distance matrix always be filled with zeros as $d(A, A) = 0$.

The minimum distance that we can find is between the taxon B and taxon C, therefore B and C form a new group (Fig. 13).

$d(B,C)=1$

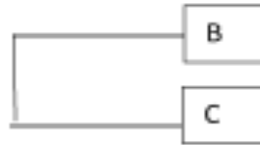


Figure 12 Conglomerate (Own Creation)

Now the selected groups into the data matrix will be modified as shown in the data matrix (Matrix 7)

Matrix 7 Data Matrix (Own Creation)

Distance Matrix			
Taxón	Characteristics		
	1	2	3
A	1	1	1
B-C	0	1	0

The new distance matrix is defined as follows: (Matrix 8):

Matrix 8 Distance Matrix (Own Creation)

Distance Matrix		
Distancia	A	B-C
A	0	
B-C	2	0

$d(A,B-C)=2$

As can be seen, the distance matrices are calculated based on the data matrix and joins the group with its nearest neighbor (Fig. 14).

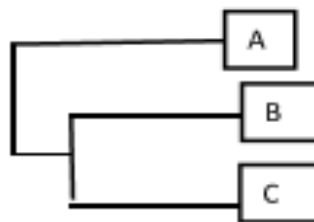


Figure 13 Conglomerate between A-B-C (Own Creation)

Creating a phylogenetic tree based on the cluster algorithm Linkage Simple

Matrix 9 is a data matrix taken from [3] which is used as a base for creating a cladogram using the Simple algorithm LinkAge

Matrix 9 Real Data Matrix [3]

Ejemplo 3	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
1 Pachynema	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
2 Pythecinus	1	1	1	1	0	1	0	0	1	1	0	0	1	1	2	1	1	0	0	1	0
3 Megaceros	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	0	1	0	1	2	1
4 Calliobrachas	1	1	1	1	1	1	0	1	0	1	0	1	0	2	1	0	1	1	2	1	1
5 Coniobrachas	1	1	1	1	1	1	1	0	0	1	0	0	0	1	1	0	1	0	1	2	1
6 Rhyparochas	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	1	1	1
7 Pectobrachas	1	1	1	1	1	1	0	1	1	1	0	0	0	1	1	1	1	1	1	1	1
8 atelivatus	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	2	1	1
9 walker	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	2	1	1
10 lumae	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	1	2	1	1

From the matrix 9 proceed to calculate the first distance matrix. We must identify into the matrix which are the taxa that have less distance to each other (minor differences). An important factor to consider in the distance matrix data is that there may be different intersections between species containing the minimum distance 0 (zero). It is for this reason that the number of trees increases based on the number of taxa involved in the analysis as specified above (Matrix 10).

Matrix 10 Distance Matrix (Own Creation)

Matriz de Distancias										
	1	2	3	4	5	6	7	8	9	10
1	0									
2	15	0								
3	11	6	0							
4	13	8	4	0						
5	12	7	4	5	0					
6	13	8	4	0	5	0				
7	14	5	5	3	4	3	0			
8	13	8	4	0	5	0	3	0		
9	13	8	4	0	5	0	3	0	0	
10	13	8	4	0	5	0	3	0	0	0

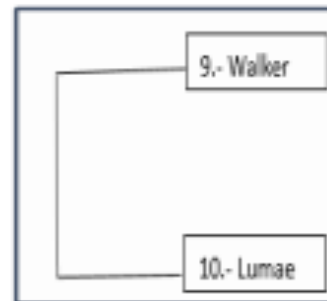


Figure 14 Conglomerate (Own Creation)

1. In the matrix 10 it can be seen that the groups have less distance each are the groups 910. Therefore these two groups have formed a new conglomerate that consists of the union of the two groups mentioned in Fig. 15 graphically describes that Union.
2. Recalculate the distance matrix (Matrix 11)

Matrix 11 Distance Matrix (Own Creation)

Matriz de Distancias									
	1	2	3	4	5	6	7	8	910
1	0								
2	15	0							
3	11	6	0						
4	13	8	4	0					
5	12	7	4	5	0				
6	13	8	4	0	5	0			
7	14	5	5	3	4	3	0		
8	13	8	4	0	5	0	3	0	
910	13	8	4	0	5	0	3	0	0

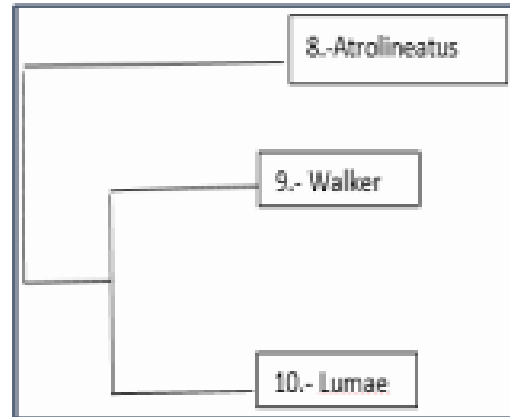


Figure 15 Conglomerate (Own Creation)

The next element to be joined is given by the group 8. In the matrix 11 one can realize that effectively the groups sharing less distance are group 8 and group 910. Therefore, they must be attached to form the group 8910 (Fig. 16).

3. Recalculate the distance matrix (Matrix 12):

Matrix 12 Distance Matrix (Own Creation)

Matriz de Distancias									
	1	2	3	4	5	6	7	8910	
1	0								
2	15	0							
3	11	6	0						
4	13	8	4	0					
5	12	7	4	5	0				
6	13	8	4	0	5	0			
7	14	5	5	3	4	3	0		
8910	13	8	4	0	5	0	3	0	

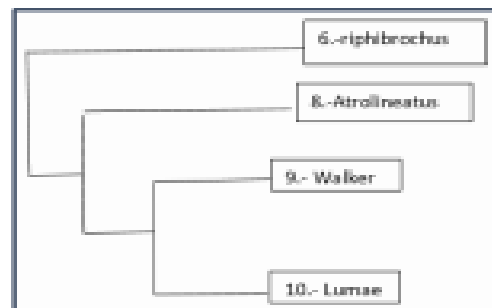


Figure 16 Conglomerate (Own Creation)

However, as we have seen, distance matrices must be calculated every time a new group is created. In the die 12 it can be seen that group 5 and intersection 8910 with the group are those with the minimum distance to each other. There are other intersections with different groups which could form new conglomerates. However, in our example the value of minimum intersection is found more in depth (see Figure 17) will be taken.

4. Recalculates the distance matrix obtained matrix 13:

Matrix 13 Distance Matrix (Own Creation)

Matriz de Distancias							
	1	2	3	4	5	7	68910
1	0						
2	15	0					
3	11	6	0				
4	13	8	4	0			
5	12	7	4	5	0		
7	14	5	5	3	4	0	
68910	13	8	4	0	5	3	0

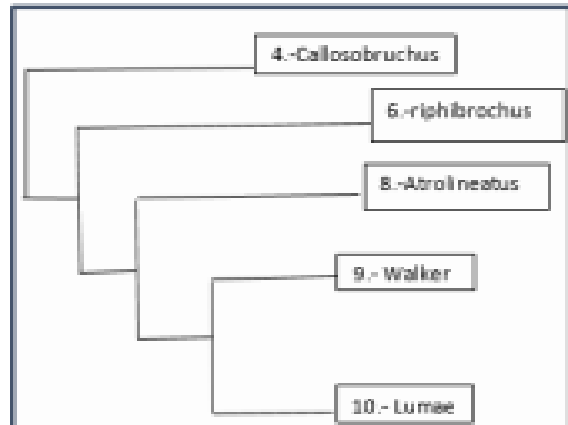


Figure 17 Conglomerate (Own Creation)

The next group to form is given by the 4 and the group 68910, thus our new group will be called 468910 which are the groups that have been formed. This representation is reflected in cluster 4 (Fig. 18).

5. Recalculating the distances, the group 7 is added to the group 468 910 because in the distance matrix (Matrix 14) can be seen which are the groups in which there is less distance. The representation can be seen in Fig. 19.

Matrix 14 Distance Matrix (Own Creation)

Matriz de Distancias						
	1	2	3	5	7	468910
1	0					
2	15	0				
3	11	6	0			
5	12	7	4	0		
7	14	5	5	4	0	
468910	13	8	4	5	3	0

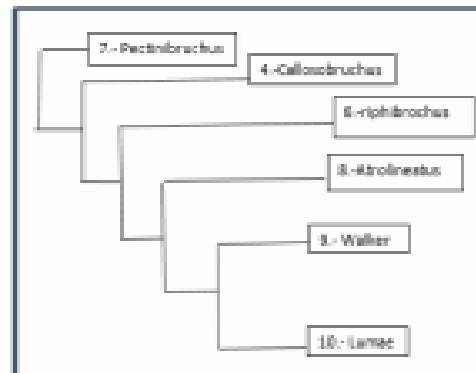


Figure 18 Conglomerate (Own Cration)

6. Proceed to do the calculations for the following matrix of distances resulting in the matrix

15. *Matriz 15 Distance Matrix (Own Creation)*

Matriz de Distancias					
	1	2	3	5	7468910
1	0				
2	15	0			
3	11	6	0		
5	12	7	4	0	
7468910	11	7	3	4	0

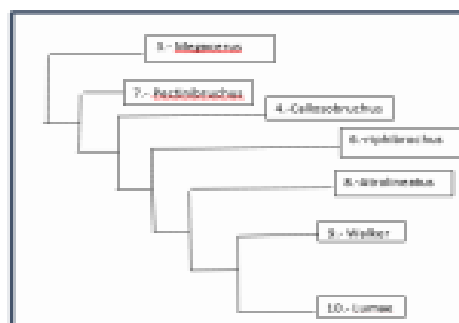


Figure 19 Conglomerate (Own Creation)

The new cluster is formed by joining the following groups with the shortest distance to each other. In the matrix of calculated distances (matrix 15) as in all previous distances matrices, we should calculate the distance matrix based on the data matrix. At this point to note that the matrix leaves the group 3 should join the conglomerate formed, their representation can be seen in the cluster (Fig. 20).

7. We calculate the distance matrix

Matriz 16 Distance Matrix (Own Creation)

Matriz de Distancias				
	1	2	5	37468910
1	0			
2	15	0		
5	12	7	0	
37468910	10	5	4	0

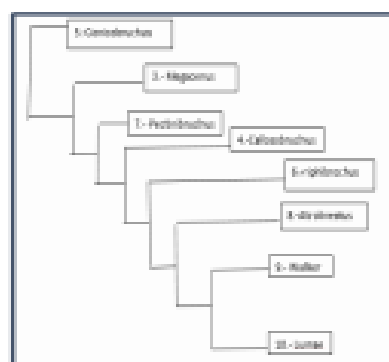


Figure 20 Conglomerate (Own Creation)

The process is repeated for each iteration adding the lower groups apart. In this iteration group 5 joins the cluster as shown in Fig. 21.

8. In conclusion, the final matrix is calculated distances to finish taxa and form unite our overall conglomerate

Matriz 17 Distance Matrix (Own Creation)

Matriz de Distancias			
	1	2	537468910
1	0		
2	15	0	
537468910	9	6	0

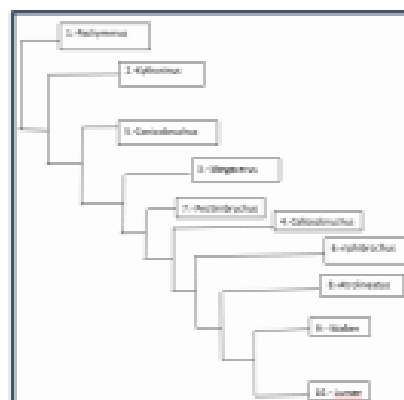


Figure 21 Conglomerate (Own Creation)

Our conglomerate to this point is almost complete. Up to this point we have the group which joins 537468910 with the Group 2 (Fig. 22).

As the distance matrix decreases, the graph increases, i.e., when a new cluster is created, we have the ability to identify which groups are those with minor differences and merge them together. When a taxon in comparison with other taxa (both belonging to the study group) found the less distance, it is said that has found his nearest neighbor.

In Fig. 23a and 23b two cladograms based on the same data matrix are shown. You can see the similarity presenting the two cladograms each other. This is, the data matrix remains the basis for the process of creating the cladograms however, and these reconstruction methods are totally different.

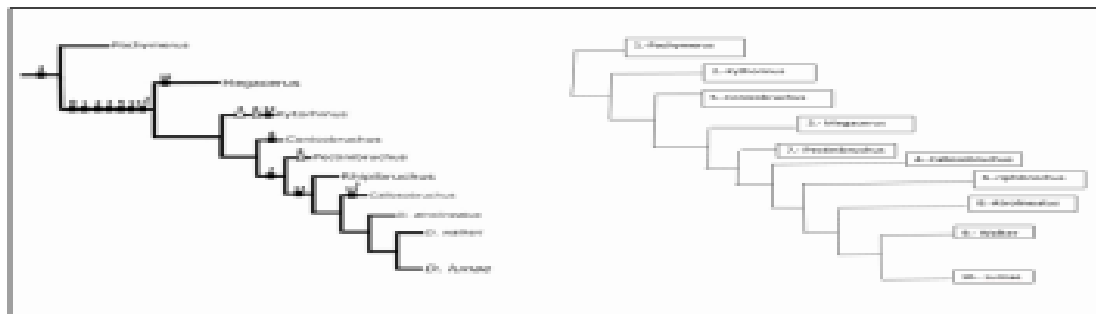


Figure 23a Phylogenetic Tree created with Hennig

Figure 23b Tree created with Clustering algorithm Simple LinkAge (Own Creation)

VII. CONCLUSIONS

Phylogenetic analysis within entomology is a process that requires validation and work by the expert. Relations between taxa should be based on the similarities that have to take a phylogenetic tree that allows us to reconstruct the similarities between organisms.

The construction of phylogenetic trees or cladograms can be carried out by different methodologies that are based on probabilistic and mathematical principles which determine each of the different forms of reconstruction. For this reason, the use of clusters within the entomological field facilitates to the expert the phylogenetic analysis because can be used as a basis tool for generating genealogical relationships between species. It is convenient to say that the final results can be very similar to each other because are based on the original data matrix which tries to link the most essential features for analysis.

Cluster analysis uses many algorithms that can be used together to deliver different results, that is why its use within the economics, statistics and many disciplines their participation is essential for representing structured data [11].

There is no standardization in formulating cladograms since it depends on the methods used as mentioned above, also the data matrices may have different characteristics because of groups of study that are to be analyzed.

In computational part, as shown in Fig. 9 and 12, although the algorithmic complexity in both algorithms has an order $O(n^3)$ shows that the algorithm is simpler cluster for its understanding and programming.

REFERENCES

- [1]. J. Cañizares, "Bioinformatica," 29 Abril 2016. [Online]. Available: <http://personales.upv.es/jcanizar/bioinformatica/filegenias.html>.
- [2]. P. Rodriguez Catalán, "Análisis Filogenéticos," Septiembre 2001. [Online]. Available: http://www.academia.edu/3578130/AN%C3%81LISIS_FILOGEN%C3%89TICOS. [Accessed 05 Septiembre 2015].
- [3]. J. R. Nápoles, "Systematics of the seed beetle genus Decellebruchus Borowiec, 1987 (Coleoptera: Bruchidae)," 2015.

- [4]. J. H. y. R. R. S. Camin, "A method for deducing branching sequences in phylogeny," 1965, pp. 311-326.
- [5]. P. Rodríguez, Análisis Filogenéticos, 2001.
- [6]. A. Tato Gomez , "Grupo de Bioinformática de la Facultad de Matemáticas," 2011. [Online]. Available: <http://mathgene.usc.es/carsoverano/cv2005/materiales/filogenia/filogenia1.pdf>. [Accessed 04 Octubre 2015].
- [7]. Introducción al análisis de Filogenias, "Bioinformatics at COMAV," [Online]. Available: bioinf.comav.upv.es/courses/intro_bioinf/filogenias.html. [Accessed 17 05 2016].
- [8]. J. M. Castillo-Cerón and I. Goyenechea, "Conceptos Básicos en Sistemática Filogenética: Los Deuterostomados como ejemplo.," in *La sistemática: base del conocimiento de la biodiversidad.* , Pachaca, UAEH, 2007, pp. 145-157.
- [9]. D. Lipscomb, *Basics of Cladistic Analysis*, Washington D. C.: George Washington University, 1998.
- [10]. S. d. I. F. Hernández, *Análisis de Conglomerados*, Madrid, España: Universidad Autónoma de Madrid, 2011.
- [11]. A. Justel, *Técnicas de análisis multivariante para agrupación*, Universidad Autónoma de Madrid.
- [12]. G. y. N. I. P. Nelson, *Systematics and biogeography: Cladistics and vicariance*, New York: Columbia University Press, 1981.
- [13]. R. Ferris, "informatica.uv.es," [Online]. Available: informatica.uv.es/iiguia/AED/oldwww/2001_02/Teoria/Tema_14.pdf. [Accessed 23 09 2015].
- [14]. *Informatica Aplicada al Análisis Económico*, Fondo Social Europeo.
- [15]. E. Triantaphyllou, *Multi-Criteria Decision Making Methods: A Comparative Study*, Louisiana, Us.A.: Springer-Science+Business Media B.V. , 2000.
- [16]. A. Herrera, *La Clasificación Numérica Y Su Aplicación En La Ecología*, Santo Domingo: Editorial Sanmenyear, 2000.



Chapingo, México., a 24 de octubre de 2016.

55/2016

BENITO SAMUEL LÓPEZ RAZO
CENTRO UNIVERSITARIO UAEM TEXCOCO
UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

Por medio de la presente tenemos el agrado de comunicarle que su trabajo: **"Sistema Web para la generación de Filogenias en base a caracteres homólogos"** ha sido seleccionado por nuestro cuerpo de árbitros y será publicado como capítulo en el libro Investigación en Economía, Vol. I., mismo que ha sido compilado por el personal Académico del Centro de Investigación en Economía y Matemáticas Aplicadas (CIEMA), de la División de Ciencias Económico-Administrativas de la Universidad Autónoma Chapingo, quienes aceptaron la publicación de la mencionada obra después de haber sido sometida a un proceso de arbitraje doble ciego y una vez que las correcciones señaladas por los dictaminadores fueron realizadas, según lo señalan los Lineamientos Editoriales del CIEMA. No omito comentarle que la publicación aparecerá el primer semestre de 2017.

Sin más por el momento, reciba nuestros cordiales saludos y lo invitamos a seguir participando activamente.

ATENTAMENTE
"ENSEÑAR LA EXPLOTACIÓN DE LA
TIERRA, NO LA DEL HOMBRE"



DR. FRANCISCO PÉREZ-SOTO
SECRETARIO TÉCNICO DEL CIEMA

Capítulo de libro.

Sistema Web para la generación de Filogenias en base a caracteres homólogos

LÓPEZ-RAZO, Benito Samuel[`], AYALA-DE LA VEGA, Joel^{``} y LUGO-ESPINOSA, Oziel^{```}

B. López, J. Ayala, O. Lugo

[`]Universidad Autónoma del estado de México, Centro Universitario Texcoco, Alumno de la maestría en Ciencias de la Computación; Av. Jardín Zumpango s/n Fracc. El Tejocote, Texcoco, Estado de México.

^{``}Universidad Autónoma del estado de México, Centro Universitario Texcoco, Profesor de tiempo completo en la Maestría en Ciencias de la Computación; Av. Jardín Zumpango s/n Fracc. El Tejocote, Texcoco, Estado de México;

^{```}Universidad Autónoma del estado de México, Centro Universitario Texcoco, Profesor Investigador de tiempo completo en la Maestría en Ciencias de la Computación; Av. Jardín Zumpango s/n Fracc. El Tejocote, Texcoco, Estado de México;

1. Introducción

La reconstrucción de relaciones ancestrales es uno de los aspectos más importantes en el estudio de la evolución de las especies dentro de la entomología, la reconstrucción se basa en características específicas y la forma en que participan para lograr este fin.

Por características entendemos patrones que pueden ser observables dentro de un grupo de especies específico. Matemáticamente se emplea una matriz de características para formar una estructura con la cual se pueda trabajar.

Las características se pueden trabajar de dos formas: podemos trabajar en base a datos representados como secuencias de ADN; o con características homologas que son representadas en la matriz de características como ceros (ausencia) y unos (presencia). Para cumplir con los objetivos de este trabajo se han tomado como base las características homologas.

Mediante el análisis filogenético puede medirse la similaridad entre un conjunto de especies, cuales son menos parecidos entre si y poder conocer cuales caracteres son los que aportan más información. Es por esta razón que las técnicas para la reconstrucción se dividen en: basados en distancias, y basados en caracteres por criterio de optimización.

Los métodos basados en caracteres, por criterios de optimización, son ampliamente utilizados dentro del campo entomológico. Sin embargo, antes de 1950, no existía una metodología específica para poder construir arboles filogenéticos en base a características. Por lo tanto, en 1950 el científico alemán Willi Hennig propuso una metodología con la cual poder reconstruir las relaciones ancestrales de forma que fuera probada y repetible para todo conjunto de características, método que hasta la fecha es muy utilizado por grupos de entomólogos.

El número de árboles filogenéticos que van a resultar dentro de un análisis filogenético puede ser calculado en base al número de características que participaran en este proceso. Es por esto que se deben crear métodos basados en heurísticas que nos ayudaran a acotar el número de posibles soluciones.

En este trabajo se presenta una aplicación web para la reconstrucción de árboles filogenéticos, que a diferencia de otras aplicaciones. Ésta es gratuita y está disponible en la red para apoyo y consulta. Se programaron dos algoritmos para el desarrollo de árboles filogenéticos: El algoritmo de Hennig (Lipscomb, 1998) (siendo éste aún un clásico en el mundo de la Entomología) y el algoritmo por conglomerados (una propuesta reciente en el cual muestra un algoritmo menos complejo de programar con resultados similares) (Lopez Razo, Ayala de la Vega, Lugo Espinoza, & Napoles Romero , 2016).

Se empleó el lenguaje Java para el back end (toda la parte algorítmica y de graficación).

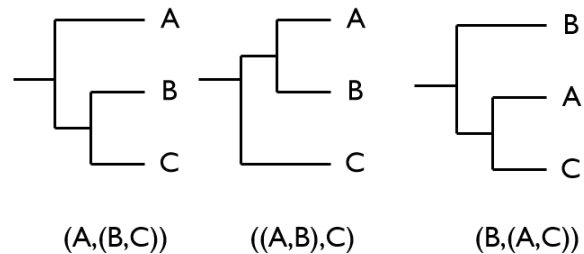
El front end fue desarrollado en lenguaje HTML y con CSS. Para la interface con Java se utilizó JavaScript. Para la recolección de los datos se empleó el lenguaje PHP.

2. Marco teórico

2.1 Intratabilidad

Como se mencionó en la introducción, el número de árboles filogenéticos depende completamente en el número de especies seleccionadas por participar. Por ejemplo, si hablamos de 3 especies (A, B, C), pueden existir 3 árboles con raíz y uno sin raíz (ver Fig. 1).

Figura 2 Árboles con 3 taxones (Tato Gomez, 2011)



Según (Rodriguez Catalán, 2001) el número posible de árboles enraizados para n taxones puede ser calculada en base a la ecuación 1:

$$N_r = \frac{(2n-3)!}{2^{n-2}(n-2)!} \quad \text{Para } n > 2$$

Donde:

Nr es la cantidad de árboles con raíz.
n es el número de taxones.

Y el número de árboles sin raíz se puede calcular mediante la ecuación 2:

$$N_u = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

Donde:

Nu es la cantidad de árboles sin raíz.
n es el número de taxones que utiliza.

El número de posibles árboles enraizados para n taxones es igual al de los árboles sin raíz para n-1 taxones. Ambos números se incrementan a medida que n aumenta. De este modo, a partir de 12 especies se vuelve difícil cuantificar el número de árboles con y sin raíz que se pudieran obtener (debido a que es un problema intratable ya que el calcular todos los posibles árboles tiene un costo computacional temporal muy elevado). Por ejemplo, un año tiene 31 536 000 segundos, un procesador Pentium IV ejecuta 4'000,000 instrucciones por segundo, por lo que aproximadamente ejecuta $126\ 144 \times 10^9$ instrucciones por año. Suponiendo que en cada instrucción se realiza un árbol, y apoyándose de la figura (ver Fig. 2), para 20 especies se tardaría 65 011 380 años en mostrar todos los árboles y para 30 especies tardaría 3.925×10^{25} años (Tato Gomez, 2011).

Figura 3 Cantidad de árboles dependiendo del número de especies (Tato Gomez , 2011)

Especies	Número de árboles
1	1
2	1
3	3
4	15
5	105
6	945
7	10.395
8	135.135
9	2.027.025
10	34.459.425
11	654.729.075
12	13.749.310.575
13	316.234.143.225
14	7.905.853.580.625
15	213.458.046.676.875
16	6.190.283.353.629.375
17	191.898.783.962.510.625
18	6.332.659.870.762.850.625
19	221.643.095.476.699.771.875
20	8.200.794.532.637.891.559.375
30	4.9518×10^{38}
40	1.00985×10^{57}
50	2.75292×10^{76}

De esta forma, cuando n es grande, el experto no puede analizar todos los árboles generados, ya que sólo uno de esos árboles representa correctamente la verdadera relación evolutiva. Por lo tanto, se utilizan heurísticas que permitan generar árboles cercanos al árbol correcto.

2.2 Hennig

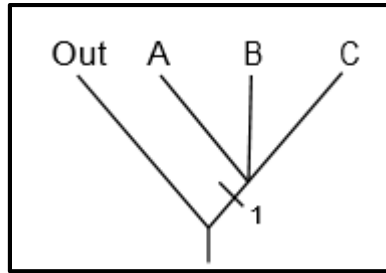
La argumentación de Hennig considera la información de cada carácter, uno a la vez. Es decir, se detecta la presencia o ausencia de dicho carácter en cada uno de los grupos seleccionados. Un ejemplo de ello se ve a continuación en la Fig. 3.

Figura 4 Matriz de Datos Hennig (Lipscomb, 1998)

	Características				
	1	2	3	4	5
<u>Outgroup</u>	0	0	0	0	0
A	1	0	0	0	1
B	1	1	0	1	0
c	1	0	1	1	0

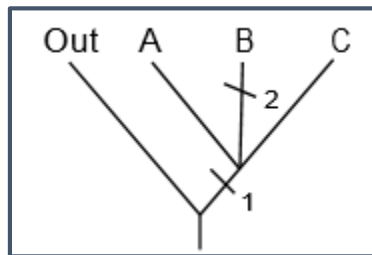
1.- El carácter 1 une los taxos (grupos) A, B y C porque ellos comparten el carácter apomorfo 1 (ver Fig. 4).

Figura 5 Árbol con el carácter 1 (Lipscomb, 1998)



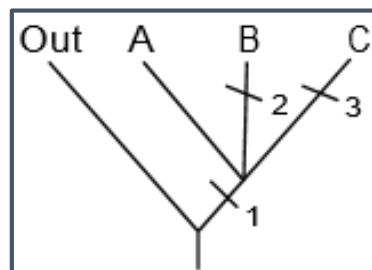
2.- Carácter 2 – el carácter derivado es encontrado solo en el taxón B, y no provee mucha información sobre las relaciones entre taxas (ver Fig. 5).

Figura 6 Árbol con el carácter 2 (Lipscomb, 1998)



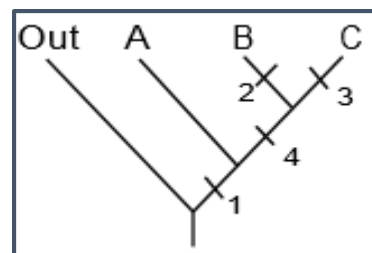
3.- Carácter 3 - el carácter derivado es autopomorfico para el grupo C (ver Fig. 6).

Figura 7 Árbol con el carácter 3 (Lipscomb, 1998)



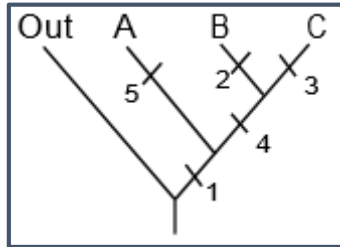
4.- Carácter 4 – el carácter derivado es sinapomorfico y une los taxas A y B (ver Fig. 7)

Figura 8 Árbol con el carácter 4 (Lipscomb, 1998)



5.- Carácter 5 – El carácter derivado es un antropomórfico para el taxón A (ver Fig. 8)

Figura 9 Árbol con el carácter 5 (Lipscomb, 1998)



Las matrices de datos reales raramente son así de simples. Sin embargo, el concepto es el mismo.

2.3 Conglomerados

El Análisis Clúster o Análisis de Conglomerados, es una técnica estadística multivariante que busca agrupar o separar elementos o variables tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos. Existen muchas técnicas para el uso de conglomerados sin embargo nos basaremos en los algoritmos jerárquicos acumulativos (forman grupos haciendo conglomerados cada vez más grandes), aunque no son los únicos posibles (Terrádez Gurrea).

Para poder unir variables o individuos es necesario tener algunas medidas numéricas que caractericen las relaciones entre las variables o los individuos. Cada medida refleja una asociación en un sentido particular y es necesario elegir una medida apropiada dependiendo del problema que se esté tratando.

Como cualquier algoritmo, es conveniente identificar los pasos que se requieren para efectuar el análisis. Los pasos dentro del análisis de conglomerados son:

4. Elección de variables
5. Elección de las medidas de asociación
6. Elección de la técnica de clúster

Dentro del análisis filogenético que a continuación se presenta, la elección de variables ha sido determinada por un experto que avala la veracidad de la matriz de datos. La medida de asociación será calculada en base a la diferencia de valores que tendrán cada una de las características homólogas con respecto al conjunto de taxones que forman la matriz de datos. Estas diferencias serán calculadas y representadas en una matriz de distancias.

Algoritmo:

- Se comienza con una matriz con n taxones (matriz de datos) y con una matriz $n \times n$ de distancias $\Delta = (\delta_{ij})$ simétrica y con ceros en la diagonal.
- Se busca en la matriz de disimilaridades los grupos que tengan menor distancia entre si (el par de grupos más próximos). Sean U y V los grupos más próximos, y d (UV) su distancia.
- Se unen los grupos U y V, y se etiqueta el nuevo grupo como (UV). Se actualiza la matriz de disimilaridades, de la siguiente forma:
 - c) se borran las filas y columnas correspondientes a los grupos U y V.
 - d) se añade una fila y una columna con las distancias entre el grupo (UV) y los grupos restantes

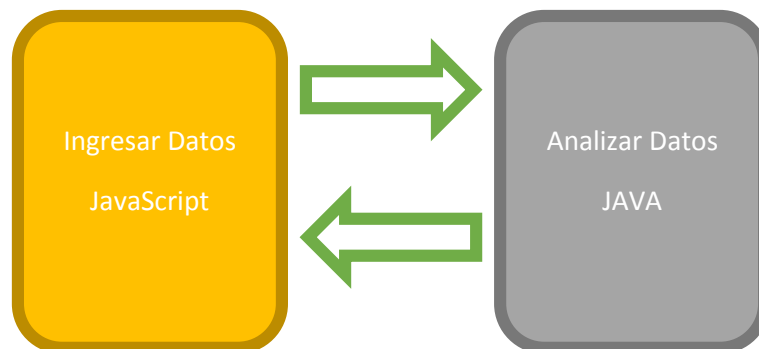
Repetir los pasos 2 y 3, $n - 1$ veces. Al final, todas las unidades estarán incluidas en un único grupo y las etiquetas de los grupos que se han unido, así como las distancias con las que se unieron (Hernández, 2011).

3.- Materiales y Métodos

La aplicación realiza el análisis de datos mediante la ejecución de un script de JavaScript instanciado por el usuario.

JavaScript Inicialmente fue desarrollado por la empresa Netscape en 1995 con el nombre de LiveScript. Posteriormente pasó a llamarse JavaScript quizás tratando de aprovechar que Java era un lenguaje de programación de gran popularidad y que un nombre similar podía hacer que el nuevo lenguaje fuera atractivo. JavaScript a diferencia de Java, se ejecuta directamente en el navegador y es por esto que nos da una forma más eficiente de manipulación de datos. En la figura 9, que se muestra a continuación, se muestran las dos principales tareas a realizar con JavaScript y la interacción entre sí.

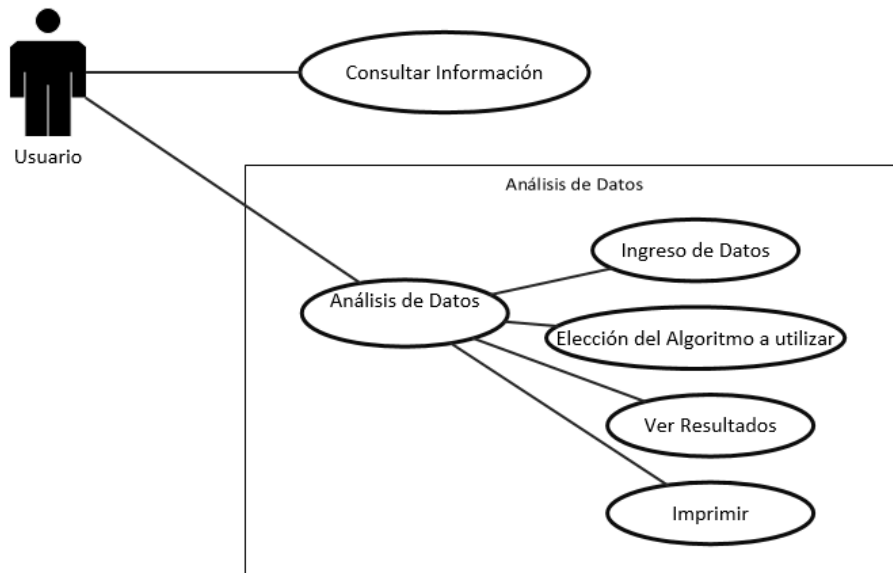
Figura 10 Uso principal de Java y JavaScript



El sistema se desarrolló utilizando la plataforma el IDE Netbeans 8.1 y se realizaron pruebas utilizando el navegador Google Chrome, sin embargo está disponible para los diferentes navegadores como Mozilla Firefox o Microsoft Edge.

La aplicación Web supone que el usuario cuenta con un grupo de estudio particular. Para poder ingresar dicha información dentro de una matriz (considerada matriz de datos). La aplicación no está acotada a un grupo determinado de especies ni a un conjunto específico de datos es por esto que los datos a analizar deben ser validados para poder obtener resultados aceptables (ver fig. 10.). Dichas acciones son observables en el siguiente Diagrama UML (Booch & Rumbaugh).

Figura 11 Diagrama UML de casos de uso



Una vez que los datos han sido introducidos, el algoritmo escogido entra en acción, en las siguientes figuras se describe el funcionamiento de cada uno. En la figura 11 se muestra el pseudocódigo del algoritmo de Hennig.

Figura 12 Seudocódigo Hennig

```

Algoritmo [sin_titulo]
+ Obtener matriz de Datos
+ Recorrer matriz de datos
+ Buscar característica con mayor incidencia
+ Copiar columna en nueva matriz
+ Terminar de recorrer matriz de datos
+ Recorrer matriz de columnas ordenadas
+ Buscar grupo (fila) con mayor número de características
+ Copiar fila a nueva matriz
+ Terminar de recorrer matriz ordenada por columnas
+ Recorrer matriz ordenada por columnas y filas
+ Identificar que grupos se unen con cada característica
+ Guardar grupos en Array
+ Terminar de recorrer matriz ordenada por columnas y filas
+ Recorrer array de grupos
+ Si array[i] = array[i+1] Entonces
+   Graficar solo una vez dicho grupo
+   Sino Graficar ambos grupos
+   Terminar de recorrer array
+ acciones por falso
Fin Si
FinAlgoritmo
    
```

Como se observa en la figura 11, el algoritmo programado requiere tres ciclos. Cada ciclo contiene anidados dos ciclos para poder recorrer la matriz $(A(n^3) + B(n^3) + C(n^3))$. Por lo que la complejidad es $O(n^3)$

Como se observa en la figura 12, Para recorrer la matriz se requieren dos ciclos anidados, por lo que el algoritmo tiene una complejidad $O(n^3)$

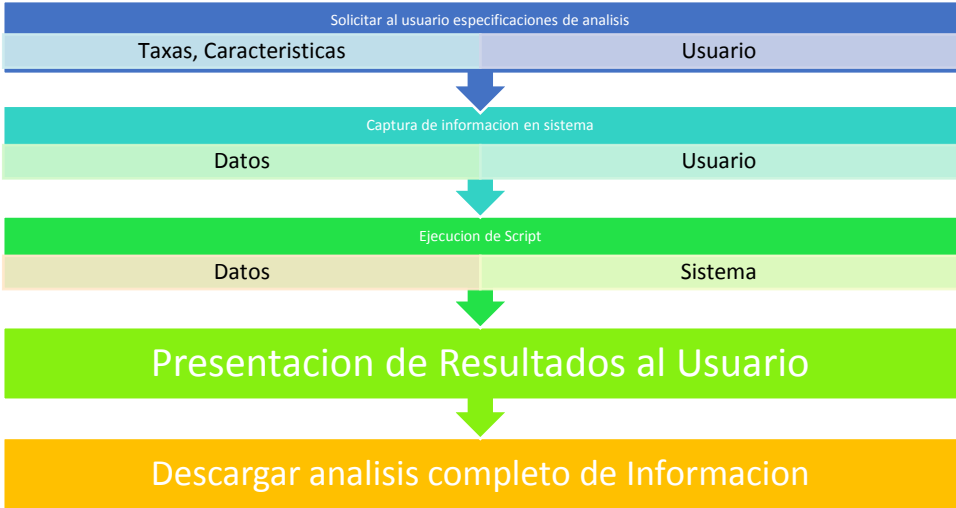
Figura 13 Seudocódigo Simple Linkage

```
Algoritmo sin titulo
  Leer matriz de datos
  Mientras expresion logica Hacer
+   Calcular matriz de distancias
+   recorrer matriz de distancias [filas][columnas]
+   min=Buscar el valor minimo()
+   if matriz de distancias [filas][columnas] es igual min
      Grupo1 = fila
      Grupo2 =columna
+   Termina if
+   Termina ciclo
+   Unix grupo1 y grupo2
+   Actualizar la matriz de datos
+   |
+   While (matriz de datos > 2)
      Fin Mientras
  FinAlgoritmo
```

Dentro del concepto de complejidad, es observable que ambos poseen una complejidad de $O(n^3)$, sin embargo son completamente diferentes al momento de ser programados, especialmente en las funciones de graficación. Simple Linkage vincula un grupo a la vez en cada iteración del análisis mientras que Hennig agrupa o separa especies en base a cada carácter que participa dentro de las iteraciones correspondientes. Esto produce que Hennig sea mucho más compleja su programación tanto en la algorítmica como en su graficación.

El funcionamiento general del sistema web está basado con la idea de que el usuario ya cuenta con la matriz de datos que se quiere trabajar. Teniendo la matriz de datos la forma en que se trabaja se muestra en la Fig. 13.

Figura 14 Funcionamiento general del Sistema WEB

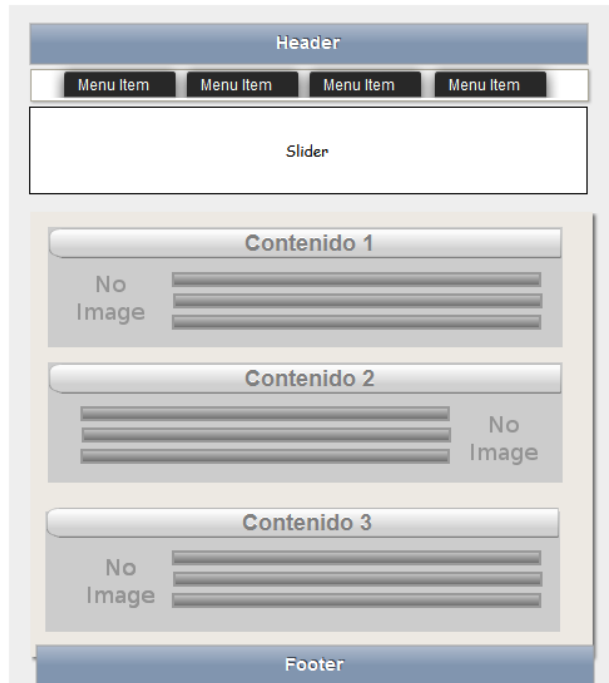


Una vez terminado el análisis, el reporte final de los datos podrá ser analizado por el usuario sin necesidad de tener que ingresar al sistema nuevamente ya que toda la información final será presentada en un documento con formato PDF.

Diseño del Sitio Web

El maquetado de la página principal del Sitio Web se muestra en la figura 14. Dentro de la página principal se busca tener una distribución óptima y un diseño que permita la facilidad en la navegación.

Figura 15 Maquetado Principal del Sitio Web



Se cuenta con una página exclusiva para poder introducir los datos de la matriz de características. El maquetado de la página para el análisis puede verse en la Figura 15. Dentro de esta página se puede indicar el número de especies, características y especificar los estados que presentan las características en cada especie.

Una vez ingresada toda la información podemos esperar a que los resultados sean desplegados y poder obtener el informe final de las características y los cladogramas.

Figura 16 Página de Análisis

The screenshot shows a web application window titled 'Hennig'. At the top right is a search bar with the text 'Buscar' and a close button. Below this are two dropdown menus: 'Taxas' and 'Características', both currently showing '10'. A 'Crear' button is positioned between these two dropdowns. Below the button is a table with three rows, each with a checkbox and a label:

<input type="checkbox"/>	Column 2
<input checked="" type="checkbox"/>	Cell Content 1
<input type="checkbox"/>	Cell content 2

Below the table is a large, empty rectangular area labeled 'Análisis de Datos'.

Conclusiones

El avance tecnológico y el acceso a medios electrónicos y digitales han incrementado la posibilidad de compartir información y conocimiento al poder desarrollar aplicaciones específicas para darle solución a un problema real. Los sitios Web son una poderosa herramienta ya que nos brindan acceso a dichos sitios Web únicamente teniendo una conexión a internet y un navegador web actualizado.

Dentro del campo entomológico, el estudio de las relaciones ancestrales entre especies es un pilar importante al momento de querer hacer un análisis de un grupo en común, y es por esta razón que la aplicación descrita anteriormente le permite al usuario reconstruir dichas relaciones de una forma sencilla y con una interfaz fácil para el usuario.

La complejidad de los algoritmos mencionados relativamente es similar ya que ambos poseen una complejidad de n^3 . Sin embargo, en la programación de los algoritmos y su graficación, simple linkage permite una programación mucho más ágil.

Referencias

- Bastar, S. G. (2012). *Metodología de la Investigación*. Estado de México : Red Tercer Milenio .
- Booch, G., & Rumbaugh, J. (s.f.). *El lenguaje Unificado de Modelado*. Recuperado el 7 de Julio de 2016, de elvex.ugr.es: <http://elvex.ugr.es/decsai/java/pdf/3E-UML.pdf>
- Goyenechea, I. (2006). Sistemática: su historia, sus métodos y sus aplicaciones en las serpientes. *Instituto de Ciencias Básicas e Ingeniería, Área Académica de Biología, Universidad Autónoma del Estado de Hidalgo*, 9. Recuperado el 20 de 11 de 2014, de <http://entomologia.rediris.es/documentos/taxonomia.htm>
- Goyenechea, I. (2006). Sistemática: su historia, sus métodos y sus aplicaciones en las serpientes del género *Conopsis*. *Instituto de Ciencias Básicas e Ingeniería, Área Académica de Biología, Universidad Autónoma del Estado de Hidalgo*.
- Hernández, S. d. (2011). *Análisis de Conglomerados*. Madrid, España: Universidad Autonoma de Madrid.
- Lara, S. L. (2015). El método que nos une: el empleo de la cladística en Antropología . *Desde el Herbario CICY* .
- Linnaeus, C. (1735). *Systema Naturae*. Sweden.
- Lipscomb, D. (1998). *Basics of Cladistic Analysis*. Washington D. C.: George Washington University.
- Lopez Razo, B. S., Ayala de la Vega, J., Lugo Espinoza, O., & Napoles Romero , J. (2016). Cluster Analisis as a methodology within Phylogenetic Systematics to Construct Phylogenetic Trees. *International Journal of Modern Engineering Research*, 15.
- Morrone, J. J. (2000). *EL lenguaje de la Cladística*. Mexico: UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO.
- Napoles, D. J. (1990). *Entomologia Sistemática*. ColPos.
- Nápoles, J. R. (2015). Systematics of the seed beetle genus *Decellebruchus* Borowiec, 1987 (Coleoptera: Bruchidae).
- Ramos, A. C. (2007). *La sistemática, base del conocimiento de la biodiversidad*. Pachuca, Centro, Hidalgo: Universidad Autónoma del Estado de Hidalgo .
- Rodriguez Baena, L. (s.f.). *www.colimbo.net*. Recuperado el 03 de Julio de 2016, de http://www.colimbo.net/documentos/documentacion/fipo/Maquetar_con_CSS.pdf
- Rodriguez Catalán, P. (Septiembre de 2001). *Anàlisis Filogenètics*. Recuperado el 05 de Septiembre de 2015, de academia.edu: http://www.academia.edu/3578130/AN%C3%81LISIS_FILOGEN%C3%89TICOS

Tato Gomez , A. (2011). *Grupo de Bioinformática de la Facultad de Matemáticas*. Recuperado el 04 de Octubre de 2015, de <http://mathgene.usc.es/cursoverano/cv2005/materiales/filogenia/filogenia1.pdf>

Terrádez Gurrea, M. (s.f.). <http://www.uoc.edu/in3/emath/docs/Cluster.pdf>. Recuperado el 07 de Julio de 2016, de <http://www.uoc.edu/in3/emath/docs/Cluster.pdf>.

Washington, U. o. (s.f.). *Phylogeny Programs*. (University of Washington) Recuperado el 14 de Julio de 2015, de <http://evolution.genetics.washington.edu/phylip/software.html>