Universidad Católica de Santa María

Facultad de Ciencias Farmacéuticas, Bioquímicas y Biotecnológicas

Escuela Profesional de Ingeniería Biotecnológica

INFERENCIA BAYESIANA PARA LA DETERMINACIÓN DE LA HISTORIA EVOLUTIVA Y RECONSTRUCCIÓN DE UN ÁRBOL FILOGENÉTICO DE PRECURSORES DE CICLÓTIDOS DE DIFERENTES FAMILIAS

Tesis presentada por el:

Bachiller:

Paula Olenska Catacora Padilla

para optar el Titulo Profesional de:

Ingeniera Biotecnóloga

Asesor:

Dr. Gómez Valdez Badhin

Arequipa – Perú

2018

UNIVERSIDAD CATOLICA SANTA MARIA
Facultad de Ciencias Farmacéuticas, Bioquímicas y Biotecnológicas
Escuela Profesional de Ingeniería Biotecnológica

Expediente N°. **16020119**

N° Trámite en Fac. **284-2016**
Fecha Recep. Fac. **05-05-2016**

FORMATO UNICO PARA TRAMITACIÓN DE TÍTULO PROFESIONAL

DE: ***CATACORA PADILLA, Paula Olenska***

PARA OPTAR EL TÍTULO PROFESIONAL DE **INGENIERO BIOTECNOLOGO**

"DETERMINACION DE UN ARBOL FILOGENETICO DE ESPECIES VEGETALES, PRESENTES EN EL PERU, QUE EXPRESEN CICLOTIDOS MEDIANTE METODOLOGIA COMPUTACIONAL"

DICTAMINADORES: *1) Blgo. Carlos Eitel Iván Paz Aliaga 2) Mgter. Roxana Bardales Álvarez*

DICTAMEN DE PLAN: Señor Decano de la Facultad de Ciencias Farmacéuticas, Bioquímicas y Biotecnológicas, en atención a su designación, el Jurado Dictaminador del Plan de Tesis informa que, hechas las observaciones y subsanadas las correcciones, consideramos se encuentra APTO para continuar con el trámite de acuerdo al Reglamento de Grados y Títulos de la Facultad

Atentamente

FIRMAS: (Devolver antes de 8 días hábiles) FECHA 13/05/16

ASESOR: ***Dr. Badhin Gómez Valdez***

DICTAMEN ASESORÍA:

Habiendo realizado el manuscrito y culminado la investigación, pongo a disposición de los jurados el borrador de tesis: DETERMINACIÓN DE LA HISTORIA EVOLUTIVA Y RECONSTRUCCIÓN DE UN ÁRBOL FILOGENÉTICO MEDIANTE INFERENCIA BAYESIANA DE CICLÓTIDOS PERTENECIENTES A LAS FAMILIA *VIOLACEAE, RUBEACEAE, FABACEAE y POACEAE*; ante su despacho.

FIRMA FECHA 28-03-18

DICTAMINADORES BORRADOR DE TESIS:

1) ***Dr. José Villanueva Salas***
2) ***Mgter. Roxana Bardales Álvarez***
3) ***Blgo. Carlos Paz Aliaga***

DICTAMEN FINAL: Señor Decano de la Facultad de Ciencias Farmacéuticas, Bioquímicas y Biotecnológicas, atendiendo a su designación como Dictaminadores del presente Borrador de Tesis sugiriendo se cambie el título a: ***"INFERENCIA BAYESANIA PARA LA DETERMINACION EVOLUTIVA Y RECONSTRUCCION DE UN ARBOL FILOGENETICO DE PRECURSORES DE CICLOTIDOS DE DIFERENTES FAMILIAS"*** y luego de hechas las observaciones y correcciones pertinentes, cumpliendo con las exigencias mínimas establecidas para un trabajo de investigación de Tesis profesional, por lo que consideramos APTO para continuar con los trámites estipulados en el Reglamento de Grados y Titulos de la Facultad.

Atentamente

FIRMA (Devolver antes de 15 días hábiles) FECHA 21/06/18

JURADOS: PRESIDENTE
VOCAL
SECRETARIO

FECHA 02/07/18 HORA 19.00 LOCAL SUM C-402

FIRMA DEL DECANO FECHA 21/06/18

Contenido

Contenido I

Índice de Figuras IV

Índice de Tablas VII

Glosario X

Dedicatoria XIV

Agradecimientos XV

Resumen XVI

Abstract XVII

introduccion XVIII

Hipótesis XX

Objetivos XXI

1. Marco Teórico 1

1.1. Ciclótidos 1

1.1.1. Generalidades 1

1.1.2. Aplicaciones 5

1.1.3. Ciclótidos en diferentes familias de plantas 7

1.1.4. Ciclótidos dentro de la biotecnología 9

1.2. Fundamentos básicos de la filogenia 9

1.2.1. Conceptos básicos 10

1.2.2. Cambios del material genético 13

1.2.3. Conceptos de homología 14

1.2.4. Teoría neutralista de la evolución molecular 15

1.3. Análisis filogenético 17

1.3.1. Alineamiento múltiple de secuencias 17

1.3.2. Modelo de evolución apropiado 20

1.3.3. Análisis Filogenética - Métodos clásicos 22

2. Metodología y detalles computacionales 28

2.1. Detalles computacionales 28

2.1.1. Hardware 28

2.1.2. Software 28

2.2. Metodología 30

3. Resultados y Discusión 34

4. Conclusiones 52

5. Recomendaciones 53

Referencias Bibliográficas 54

Apéndices 66

I Apéndices **66**

.1. Anexo 1 67

.2. Anexo 2 87

Índice de Figuras

1.1. Secuencia y estructura de la kalata B1. En la izquierda la estructura a través de la RMN (PDB ID: 1NB1). Al lado derecho la secuencia aminoacídica con nudo cíclico de cisteína. Fuente: Craik, David J(2015) [1] 2

1.2. Linea de tiempo sobre los hallazgos e investigaciones de los ciclótidos. Fuente: Craik, David J(2015) [1] 3

1.3. Representación del precursor de un ciclótido.Fuente Gunasekera (2006)[2] . 4

1.4. Descripción general de las aplicaciones agrícolas y del potencial farmacéutico. Fuente: Craik, David J (2010)[3]. 7

1.5. Terminología del árbol filogenético.Identificando el nodo, la rama, la topología, longitud de la rama, raíz y la escala entre distancias. Fuente: Wilkinson and Mark and McInerney (2007)[4]. 10

1.6. Árbol no enraizado (a) y sus siete posibles opciones de árboles enraizados (b). Fuente: Wilkinson and Mark and McInerney (2007)[4]. 12

1.7. Ilustración del concepto de genes ortólogos y parálogos. Genes A y B fueron derivados por duplicación y cada uno sufrió un evento de especiación. Fuente: Bork, Peer and Dandekar (1998)[5]. 16

1.8. Logaritmo de MUSCLE que consta de tres etapas: primera etapa de alineación progresiva, segunda etapa de optimización del método progresivo y tercera etapa de reimplementación. Fuente: Edgar, Robert C (2004)[6]. . 18

1.9. (A) la representación hipotética de secuencias; Y sus diferentes topologías (B) Parsimonia cuando los gaps son usados caracteres; (C) Parsimonia cuando los gaps son tratados como información faltante; (D) topología hecha por análisis de maximum likelihood. Fuente: Giribet, Gonzalo and Wheeler, Ward C (1999)[7]. 19

1.10. Diagrama de flujo del funcionamiento del programa JModelTest. Fuente: David Posada(2008) [8]. 21

1.11. Diagrama de flujo básico del funcionamiento del programa ProtTest. Fuente: Federico Abascal, Rafael Zardoya and David Posada (2005)[9]. 22

1.12. (a)Un árbol semejante a una estrella sin estructura jerárquica Y (b), un árbol en el que las UTO 1 y 2 están agrupadas. Fuente: Naruya Saitou and Masatoshi Nei (1987)[10]. 24

2.1. Obtención de las secuencias aminoacídicas de los ciclótidos en el servidor Cybase hasta febrero del 2017 31

2.2. Secuencias alineadas mediante el programa UGENE 32

2.3. Comandos utilizados con el programa DAMBE 32

3.1. Prueba de saturación. Los triángulos en verde muestran las transversiones y las cruces en azul las transiciones. Eje de las X muestra la distancia genética corregida F84; el eje de las Y muestran el número de substituciones nucleotídicas. 40

3.2. Estimación a porsteriori del LnL (logaritmo natural de verosimilitud) realizada por las dos corridas en conjunto mostrando una distribución normal, que significa stacionariedad y convergencia. 42

3.3. Estimación a porsteriori del LnPr (logaritmo natural de probabilidad) realizada por las dos corridas en conjunto mostrando una distribución normal, que significa stacionariedad y convergencia. 43

3.4. Gráfica de la probabilidad de distribución marginal de la transición de los nucleótidos mediante en parámetro KDE (kernel density estimate). . . . 44

3.5. Gráfica de la probabilidad de distribución marginal de la transición de los nucleótidos mediante un histograma, para observar el ruido del análisis. . 45

3.6. Gráfico tipo tracer comparando las cadenas MCMC de la primera corrida en el eje x contra los valores de verosimilitud logarítmica (LnL) en el eje y. 46

3.7. Gráfico tipo tracer comparando las cadenas MCMC de la segunda corrida en el eje x contra los valores de verosimilitud logarítmica (LnL) en el eje y. 46

3.8. Gráfico tipo tracer comparando las cadenas MCMC de las corridas en conjunto en el eje x contra los valores de verosimilitud logarítmica (LnL) en el eje y. 47

3.9. Árbol filogenetico realizado con inferencia bayesiana, analizando los datos de forma separada solo co. N de generaciones =25000000, burnin = 25 % 49

3.10. Árbol filogenetico realizado con inferencia bayesiana, analizando los valores superiores e inferiores dentro de la densidad de probabilidad más alta al 95 % 50

3.11. Estimación de tiempo de divergencia en millones de años sin calibración de fósiles usando un reloj molecular relajado 51

Índice de Tablas

3.1. Resultado de la depuración bibliográfica de los ciclótidos a usar en el estudio filogenético; con su respectiva familia, especie, número de acceso y número de nucleótidos 35

3.2. Recopilación de la información del Grupo externo(Outgroup) a usar en el estudio filogenético 36

3.3. Leyenda de Ciclótidos a usar en la Tabla 3.4 37

3.4. Resultado de la similitud en porcentajes comparando los 21 ciclótidos y el outgroup usando el programa UGENE 38

3.5. Resultados obtenidos mediante el programa JModelTest para la elección del modelo de sustitución nucleotídica para cada tipo de criterios 39

3.6. Resultado de los primeros 5 modelos de sustitución nucleotídica para el riterio BIC 40

3.7. Parámetros estimados con el programa Tracer con sus respectivas medias y tamaño de muestra efectiva (ESS). 42

3.8. Resumen estadístico para el parametro LnL. 43

3.9. Resumen estadístico para el parametro LnPr. 43

3.10. Tabla sobre los saltos y burn-in de las cadenas en separado y en conjunto respectivamente. 45

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 68

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 69

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 70

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 71

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 72

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 73

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 74

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 75

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 76

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 77

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 78

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 79

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 80

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 81

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 82

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 83

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 84

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 85

1. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos 86

2. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de acceso y número de aminoácidos 88

2. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de acceso y número de aminoácidos 89

2. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de acceso y número de aminoácidos 90

2. Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de acceso y número de aminoácidos 91

Glosario

1. A = Adenina
2. ACT = autocorrrelaciónde tiempo, promedios de estado en la cadena MCMC para que dos muestras sean independientes de la posterior.
3. AIC = Criterio de información de Akaike
4. AICC = Criterio de información de Akaike corregido
5. Algoritmo de MH = Algoritmo de Metropolos Hastings
6. Analogía = Caracteres parecidos pero no homólogos, realizan las mismas funciones biológicas. Puede ser convergencia o paralelismo
7. IB = Inferencia bayesiana
8. BIC = Criterio de información Bayesiana
9. Bifurcación = Un nodo en un árbol que conecta exactamente tres ramas. Si el árbol es dirigido (rooted), entonces una de las ramas representa un linaje ancestral y las otras dos ramas representan los linajes descendientes.
10. BLASTA = (Basic Local Alignment Search Tool) es un programa informático de alineamiento de secuencias de tipo local, ya sea de ADN, ARN o de proteínas.
11. Burnin = Arboles descartados como parte del proceso de todas las muestras.
12. C = Citosina
13. Carácter = Un rasgo que es una parte observable o un atributo de un organismo (puede ser anatómico, etológico, genómico, bioquímico...)

14. CCK = Nudo de cisteína enlazado por puentes de disulfuro
15. CI = Índice de consistencia: Recíproco del número de veces que aparece el rasgo en el árbol. "
16. Clado = Un grupo monofilético en un cladograma. Los clados no aparecen los fenogramas, árboles de distancia..., son exclusivos de los cladogramas.
17. Convergencia = Se da cuando dos estructuras similares han evolucionado independientemente a partir de estructuras ancestrales distintas y por procesos de desarrollo muy diferentes
18. Criterio de parsimonia = Se prefiere el árbol que implique el menor número de eventos para explicar la distribución observada de los rasgos entre los taxa
19. Especiación = Aparición de diferencias entre dos especies próximas, que motiva su separación definitiva.
20. et al. = Del latín .et alii"que significa ƨ otros".
21. Divergencia = Bifurcación de una rama padre en dos ramas hijas en cualquier punto dado.
22. ESS = tamaño de muestra efectiva
23. FASTA = Es un formato de fichero informático basado en texto, utilizado para representar secuencias bien de ácidos nucleicos, bien de péptido, y en el que los pares de bases o los aminoácidos se representan usando códigos de una única letra.
24. Fmoc = Fluorenil-9-metoxicarbonil, tipo de grupo de aminoácido.
25. G = Guanina
26. Genes ortólogos = Genes que presentan homología en dos especies distintas
27. Genes parálogos = Dos genes de una misma especie que han evolucionado por duplicación génica.
28. Grupo externo (outgroup) = Es cualquier grupo usado en el análisis que no es incluido en el taxón bajo estudio. Se utiliza para fines comparativos y debe ser lo más cercano posible al grupo interno, preferentemente su grupo hermano.

29. Grupo interno (ingroup) = Es el grupo actualmente estudiado por el investigador.

30. Heurística = Es un algoritmo que abandona uno o ambos objetivos; por ejemplo, normalmente encuentran buenas soluciones, aunque no hay pruebas de que la solución no pueda ser arbitrariamente errónea en algunos casos; o se ejecuta razonablemente rápido, aunque no existe tampoco prueba de que siempre será así.

31. HPD = 95 % de dlas densidades posteriores, intervalo de confianza bayesiano.

32. MCMC = Cadenas markovianas de Monte Carlo.

33. UTO = Unidad taxonómica operativa, es una unidad de clasificación seleccionada por el investigador que la utiliza para individualizar a objetos de su estudio, ya sea una especie.

34. pb = Pares de bases.

35. Pr(D) = Probabilidad incondicional de los D, que puede ser obtenida usando la ley de la probabilidad total.

36. Pr(D|H) = Probabilidad posterior; probabilidad de H (o valor del parámetro), dados D

37. Pr(H) = Probabilidad anterior; es la prob. incondicional de H

38. Prior = Probabilidad anterior.

39. RMN = Resonancia magnética nuclear.

40. SNP = Un polimorfismo de un solo nucleótido, una variación en la secuencia de ADN que afecta a una sola base (adenina (A), timina (T), citosina (C) o guanina (G) de una secuencia del genoma.

41. Verosimilitud = Es la credibilidad o congruencia de un elemento determinado dentro de una obra de creación concreta.

42. T = Timina.

43. Transición = Cuando se sustituye una purina por purina (A <–>G) o una pirimidina por pirimidina (C <–>T)

44. Transversiones = Cuando se sustituye una pirimidina por una purina o viceversa (T o C <–>G o A).

Dedicatoria

Dedicado a mi mamá, gracias por absolutamente todo

Agradecimientos

Esta tesis es la culminación de mi etapa Universitaria y no puedo deja de agradecer a mi mamá por su apoyo y cariño incondicional, sin ti nada sería posible, a mi familia en general mi papá, mi abuelo, mis abuelas y mi primo gracias por estar siempre ahí; al CIIM y a mi asesor gracias por la confianza brindada y por guiarme al camino de la investigación, finalmente a mis amigos del centro de investigación en ingenieria molecular, fue un camino largo y dificultoso, pero con su compañía hasta fue divertido.

Resumen

En los últimos años, se ha presentado un gran interés en el estudio de especies que contienen ciclótidos entre sus metabolitos secundarios, los ciclótidos son oligopéptidos cíclicos con un nudo de cisteína por efecto de la presencia de tres puentes disulfuro, presentado diversas propiedades antitumoral, anti-VIH, etc. Sin embargo, su distribución filogenética aún no está claramente definida, teniendo en cuenta los avances tecnológicos y computacionales, debería ser factible el análisis filogenético. El objetivo de este estudio fue: estudiar a los ciclótidos a nivel filogenético usando paquetes computacionales y su conjunto de algoritmos aplicados a la biología computacional, con el fin de profundizar las relaciones evolutivas de ciertos organismos que los expresan, para poder analizar y relacionar estudios por familia. Para ello se analizaron 20 secuencias de mRNA precursor de ciclótidos pertenecientes a las familias *violaceae, rubiaceae, fabaceae* y *poaceae*, usando como grupo externo el ciclótido hipotético Hcf-1, con la característica de ser una secuencia precursora derivada de estudios anteriores, con esta información, se generó una matriz de secuencias alineadas que fue utilizada para llevar a cabo la reconstrucción filogenética mediante análisis bayesiano. Las topologías de los árboles resultantes permitió determinar que los ciclótidos parecen tener altas similitudes en su secuencia y en su identidad estructural; este alto grado de homología sugiere la conservación durante su evolución en las plantas. El estudio sobre las familias *rubiaceae* y *violaceae* nos permite concluir que podría haber existido una proteína ancestral común antes de la separación de estas, siendo familias de plantas lejanamente relacionadas. Finalmente nos permitió concluir que es posible que las proteínas que están relacionadas filogenéticamente compartan una

función común aún no probada. Por lo tanto, es necesario realizar más estudios de la relación de función estructural y estos también pueden mejorar nuestro conocimiento sobre la evolución funcional.

Palabras clave: Análisis filogenético, inferencia Bayesiana, ciclótidos, biología computacional, evolución functional.

Abstract

In recent years, there has been great interest in the study of cyclotides containing cells in their secondary metabolites, cyclotides are cyclic oligopeptides with a cysteine knot due to the presence of three disulfide bridges, have several properties like antitumor, anti-HIV, antibacterial and anti insecticide. However, its phylogenetic distribution is not yet clearly defined, taking into account technological and computational advances, phylogenetic analysis should be feasible. The objective of this study was: to study the cyclotides at a phylogenetic level using computational packages and their set of algorithms applied to computational biology, in order to deepen the evolutionary relationships of certain organisms that express them, in order to analyze and relate studies by family. For this it, 20 precursor mRNA sequences of cyclotides belonging to the families textsl Violaceae, Rubeaceae, Fabaceae and textsl Poaceae were analyzed, using as an external group the cyclotide hypothetical Hcf-1, being this a derived precursor sequence from previous studies. With this information, a matrix of aligned sequences was generated that was used to carry out the phylogenetic reconstruction of Bayesian analysis. The topologies of the resulting trees that determine that the cyclotides seem to have similarities in their sequence and in their structural identity; this high degree of homology suggests conservation during its evolution in plants. The study on the families Rubiaceae and Violaceae allows us to conclude that a common ancestral protein could exist before the separation of these, being distantly related families of plants. Finally, it allowed us to conclude that it is possible that proteins that are phylogenetically related share a common function not

yet proven. Therefore, it is necessary to carry out more studies of the structural function relationship and these can also improve our knowledge about functional evolution.

Keywords: Phylogenetic analysis, Bayesian inference, cyclotides, computational biology, functional evolution.

Introducción

El Perú consta de una diversidad biológica muy amplia, tanto en fauna como en flora, dentro de la flora endémica se hallan un número notable de plantas con propiedades medicinales, un extenso rango de estas pertenecen a las familias *Violaceae* Violaceae y *Rubiaceae*; la evidencia presume que uno de los factores por el cual presentan esta cualidad es la presencia de ciclótidos dentro de sus metabolitos secundarios. Los ciclótidos son oligopéptidos cíclicos con un nudo de cisteínas por efecto de la presencia de tres puentes disulfuro, presentando así diversas propiedades biológicas, farmacológicas y como insecticidas siendo así las más destacadas antitumoral y anti-VIH; en la actualidad hay un auge de investigaciones en ciclótidos dentro de la comunidad científica, llevándonos a descubrir cada día nuevos ciclótidos, funciones, familias, etc., las nuevas familias donde se han revelado la existencia de ciclótidos son las de *Fabaceae, Solanaceae, Poaceae* y *Curcubitaceae* ; sin embargo su distribución filogenética aún no esta claramente definida.

A través de los años se ha compartido la idea de que todos los organismos están genéticamente relacionados, la filogenia es un campo dentro de la biología, cuyo término fue propuesto por el alemán Ernst Haeckel en 1866, que trata de la búsqueda de las relaciones e historia evolutiva del linaje de los organismos; estas relaciones genéticas son representadas por un árbol evolutivo llamado el árbol de la vida.

Charles Darwin fue el escritor del origen de las especies, donde formuló las bases científicas sobre la teoría de la evolución por medio de la selección natural; actualmente un análisis filogenético no solo nos indica las relaciones evolutivas entre las secuencias o especies, cuales descienden de ancestros comunes, también puede indicarnos cuales son las distancias entre ellas; no obstante, este tipo de estudios a variado circunstancialmente en los últimos años ya que antiguamente se usaban datos morfológicos, la desventaja es que son poco informativos, generan ruido y las investigaciones pueden tardar años, actualmente gracias a los avances computacionales se pueden usar datos moleculares, dando

paso a la filogenética computacional. La filogenética computacional tiene como objetivo fundamental trazar la relación ancestro descendiente construyendo así un árbol filogenético, con el desarrollo nuevas de bases de datos, métodos de simulación de secuencias y rigurosos métodos de filogenética estadística, tanto frecuentistas y Bayesianos, permiten realizar una hipótesis evolutiva verídica y confiable. Uno de los debates más interesantes entre los evolucionistas del siglo XXI ha sido por el enfoque dado a la teoría neutralista de la evolución molecular, esta teoría sostiene que la mayoría de cambios a nivel molecular es neutro y propone como mecanismo evolutivo a la deriva génica, esto permite establecer un reloj molecular y cuantificar el tiempo de divergencia; la controversia surgió cuando a finales de 1960 el japones Motoo Kimura propuso la teoría para explicar que el azar juega un papel importante dentro de la evolución molecular, sin embargo esta teoría no niega la intervención de la selección natural a nivel molecular.

La siguiente investigación tiene como propósito estudiar a los ciclótidos a nivel filogenético usando paquetes computacionales y su conjunto de algoritmos aplicados a la biología computacional, con el fin de profundizar las relaciones evolutivas de ciertos organismos que los expresan, para poder analizar y sinergizar estudios por familia.

Hipótesis

Dedo al amplio interés suscitado por los ciclótidos en la comunidad científica por las potenciales aplicaciones debido a sus propiedades, es posible que mediante el uso de inferencia bayesiana podamos definir más claramente su distribución filogenética y aclarar el panorama de su historia evolutiva.

Objetivo General

Determinar de la historia evolutiva y reconstruir un árbol filogenético a través de la inferencia Bayesiana para los precursores de ciclótidos de las familias *violaceae, rubeaceae, fabaceae* y *poaceae*.

Objetivos Específicos

1. Seleccionar las secuencias a análisar de precursores ciclotídicos y comparar las similitudes existentes a través de un alineamiento múltiple.
2. Validar la continuidad del análisis filogenético a través del indice de saturación y la elección del modelo de evolución apropiado.
3. Evaluar la estacionariedad y convergencia de las cadenas MCMC a través del tiempo mediante inferencia bayesiana.
4. Reconstruir un árbol filogenético usando el método de inferencia bayesiana para los precursores de ciclótidos de las familias *violaceae, rubeaceae, fabaceae* y *poaceae*. .

Capítulo 1

Marco Teórico

1.1. Ciclótidos

1.1.1. Generalidades

En los últimos años, se ha presentado un gran interés en el estudio de especies que contienen ciclótidos dentro de sus metabolitos secundarios, los ciclótidos son oligopéptidos pertenecientes a las plantas cuyo primer reporte pertenece a principios de la década de 1970[11], sin embargo el nombre "ciclótido" fue formalmente introducido en el año 1999 para describir a proteínas con la característica distintiva de ser un péptido cíclico de la cabeza a la cola y una disposición de tres puentes disulfuro. Estos oligopéptidos contienen aproximadamente 30 aminoácidos, entre ellos 6 residuos de cisteína absolutamente conservados que forman un nudo cíclico de cisteína en el núcleo de sus estructuras,como podemos observar en la Figura 1.1[1].

Los antecedentes presentados en dos lugares diferentes de África nos dieron a conocer que en la República Centro-africana se usaba el extracto de una planta nativa llamada "Wetegere", nombre que significaba de fácil labor, este fue el primer informe oficial sobre la actividad uterotómica de la planta Oldenlandia affinis. El siguiente informe nos lleva hasta el Congo durante su periodo de conflicto en 1960 cuando el equipo de la cruz roja que trabajaba en el hospital Central de Luluabourg usaba un brebaje conocido como Kalata-Kalata con las mujeres en el momento del parto ya que este tenia efectos uterotómicos; años después el agente activo fue aislado dándole el nombre de Kalata B1[12] el

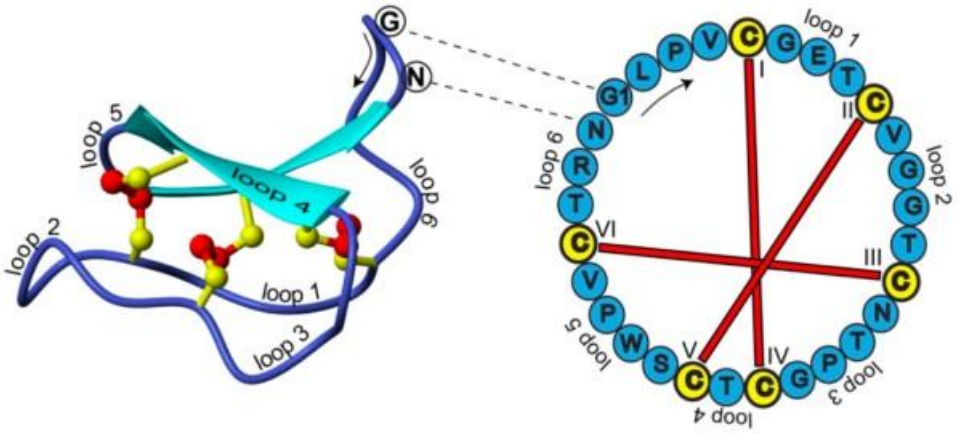


Figura 1.1: Secuencia y estructura de la kalata B1. En la izquierda la estructura a través de la RMN (PDB ID: 1NB1). Al lado derecho la secuencia aminoacídica con nudo cíclico de cisteína. Fuente: Craik, David J(2015) [1]

cual consistía de 29 aminoácidos incluyendo 6 secuencias perfectamente conservadas de cisteína. Posteriormente según los estudios de escisión enzimática se pudo demostrar que el esqueleto del oligopéptido es cíclico, la estructura tridimensional de la solución se ha determino utilizando espectroscopia de resonancia magnética nuclear (RMN) bidimensional [13].

Hoy en día, es evidente que los ciclótidos son mucho más numerosos que lo anticipado y se ha sugerido que pueden superar las conocidas defensas vegetales en número y diversidad; la gran mayoría de ciclótidos fueron aislados principalmente dentro de la familia de las violetas (Vioalecea) y la familia del café (Rubiaceae), sin embargo se han encontrado en menor proporción dentro de las familias Curcubitaceae, Fabaceae, Solanaceae y Poaceae este último de gran importancia ya que engloba los cultivos de cereales más importantes como el trigo y el maíz [14]. Los ciclótidos son oligopéptidos cíclicos que se expresan en los metabolitos secundario de las plantas, debido a su nudo cíclico de cisteína los, ciclótidos tienen una resaltante estabilidad contra la proteasa, agentes térmicos y químicos; estas características son las que los diferencian de los péptidos no cíclicos. También presentan diversas propiedades biológicas, incluyendo anti-VIH, efectos antimicrobianos, citotóxicas y diversas propiedades farmacológicas comprobadas [15].

En total existen 3 tipos de subfamilias de ciclótidos denominados Mobius, Bracelet y finalmente Inhibidor de tripsina; primeramente se encuentran los Bracelet, que son aquellos que tienen un mayor número de residuos cargados positivamente y cuentan la

Figura 1.2: Linea de tiempo sobre los hallazgos e investigaciones de los ciclótidos. Fuente: Craik, David J(2015) [1]

abstinencia de la secuencia cis prolina; otro subtipo son los llamados mobius y cuentan con la secuencia cis prolina conservada [16]; finalmente esta la familia del inhibidor de tripsina y comprende un aproximado de 34 aminoácidos siendo este la familia con mayor longitud de aminoácidos [17].

En la actualidad nuevas secuencias de ciclótidos van incrementando al igual que sus estudios(Figura 1.2), de tal manera que en el presente cuentan con una base de datos exclusivamente para ellos; CyBase (http://cybase.org.au) es una base de datos dedicada al estudio de las secuencias y estructuras tridimensionales de proteínas cíclicas y sus variantes. Actualmente consta de un catálogo muy diverso que incluye los inhibidores de tripsina, proteínas bacterianas, toxinas de hongos, ciclótidos y defensinas cíclicas. CyBase tiene herramientas especificas para los ciclótidos incluyendo representaciones bidimensionales de dominios y presentaciones alternativas de alineación para secuencias precursoras [18].

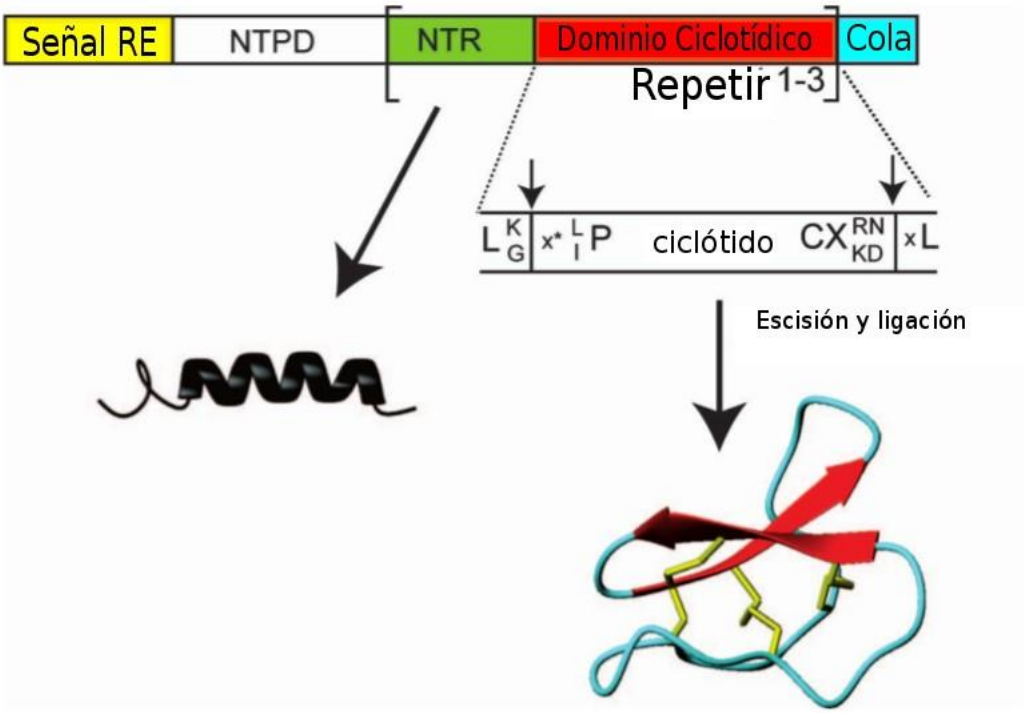


Figura 1.3: Representación del precursor de un ciclótido.Fuente Gunasekera (2006)[2]

1.1.1.1. Biosíntesis de ciclótidos

Los ciclótidos son una nueva familia de péptidos derivados de plantas que se producen de a través de biosíntesis de sus la productos génicos codificados, que derivan de la transformación de proteínas precursoras lineales que contienen de uno a tres dominios de ciclótidos [19].

Las circulares sintetizadas que son ribosomales, osea se sintetizan ahí, se codifican como genes transcrito en ARN y se sometió a procesamiento ribosomal para producir el precursor lineal representado en la Figura 1.3; la estructura genérica de las estructuras precursoras incluye un pró-domínio N-terminal (NTPD), en esta parte los genes de la misma especie están relativamente bien conservados, cada dominio consta de 28 a 37 residuos y están separados por una región de 25 aminoácidos a la cual se le llama N-terminal (NTR) también existe la secuencia de señal RE que sugiere el rompimiento y el momento de pliegue no se sabe si los NTPD y / o las RNPs tienen un papel en este proceso[2].

Las enzimas de procesamiento vacuolar son cisteína proteinasa, estas son conocidas

como VPEs y el requerimiento básico es que el pH de la vacuola sea bajo para así poder activarla [20], se cree que estas VPEs son de gran ayuda en el momento de la ciclación de los ciclótidos y su localización subcelular [21].

Para demostrar la acción de de la actividad de las enzimas de procesamiento vacuolar (VPEs), en los estudios se demostró que que es responsable de la hidrólisis del enlace asparginil dentro de los ciclótidos y también fue capaz de escindir sustratos sintéticos del ciclótido dentro del enlace asparaginil C-terminal del dominio del ciclótido, y la muerte de la actividad VPE se relaciono con un aumento de ciclótidos lineales [22].

1.1.1.2. Síntesis química de los ciclótidos

La síntesis en las proteínas circulares ahora es fácilmente alcanzable gracias a que la nueva metodología lo permite, gran parte por los avances en la ligadura química entrópica; y así para superar la barrera de la entropía en el acoplamiento de los extremos N- y C-terminales de grandes segmentos peptídicos, ya sea para la ligación intermolecular o la ligadura intramolecular en la ciclación de la estructura [23].

La ligadura intramolecular de péptidos α-ésteres implica la síntesis de péptidos en fase sólida Fmoc en el momento de la ciclación, llamada ligación por química nativa, seguido de la oxidación de replegamiento para producir la proteína de forma natural. Esta reacción secundaria correspondió a una esterificación intramolecular de la reacción del éster de metilo α-hidroxi. Mediante el uso de este nuevo enlazador se informó de la primera síntesis de la Kalata b1 nativa a través de la ligadura intramolecular del péptido oxiéster seguido de plegamiento oxidativo [24]. este proceso intramolecular no requiere de la presencia de un cofactor tiol, opositoriamente al proceso de ligadura intermolecular en el cual es absolutamente necesario [25] [26].

1.1.2. Aplicaciones

1.1.2.1. Potencial farmacológico

Usar péptidos de un tamaño muy reducido tienden a ser sumamente beneficiosos; ya que gracias a esto tienen la ventaja de ser muy específicos para sus objetivos, los péptidos son producto de miles de años de selectividad evolutiva que han sido programados para

interactuar específicamente con dianas biológicas, evolucionando en potentes hormonas endógenas, factores de crecimiento, neurotransmisores y moléculas de señalización, así como agentes inmunológicos y de defensa. Al tener tan buena afinidad para interactuar con diferentes agentes biológicos su gran desventaja y por la cual su sintetización química ha sido severamente limitada; es su su baja estabilidad sistémica, pobre permeabilidad de la membrana, actividad despreciable cuando se administra por vía oral, y sus altos costos de fabricación. A pesar de estas desventajas un análisis del mercado en el año 2013 nos indica que existen más de 100 fármacos basados en péptidos lo que se estima en unos 40.000 millones de dolares al año , es decir el 10 % del mercado farmacéutico ético; el cambio de paradigma en el interés por la industria farmacéutica de péptidos se ejemplifica en las tasas del mercado que van incrementando alrededor de todo el mundo [27]. El uso de péptidos lineales como fármacos es limitada debido a su susceptibilidad a la escisión proteolítica y la escasa biodisponibilidad; tales limitaciones pueden ser superadas por péptidos cíclicos, un ejemplo claro de esto son los ciclótidos cuya estructura simula un andamio al cual se pueden diseñar nuevas actividades gracias a la aparición de nuevos enfoques en la ingeniería de proteínas [28].

Los ciclótidos tienen una gran gama de actividades biológicas que son de gran interés terapéutico, la actividad biológica más estudiada es la de anti-VIH [29] tres ciclótidos encontrados en la planta V. Odorata en la cual la correlación entre el aumento de la hidrofobicidad y la actividad anti-VIH, [30] actúa como pétidos activos por vía oral contra el dolor inflamatorio incluyendo el cáncer y la artritis reumatoide [31], actúa como agente anti-angiogénicos, con el objetivo de reducir el crecimiento de vasos sanguíneos en los tumores [32].

1.1.2.2. Aplicaciones en la agricultura

Una de las más grandes aplicaciones de los ciclótidos, gracias a su función antimicrobiana, es como insecticida; un claro ejemplo es su potente efecto inhibidor sobre el crecimiento y desarrollo de las larvas en las plantas perteneciente al genero Helicoverpa [33].

Como se puede observar en la Figura 1.4; para las aplicaciones farmacéuticas, implica el injerto de epítopos de péptidos biológicamente activo dentro de la estructura de los

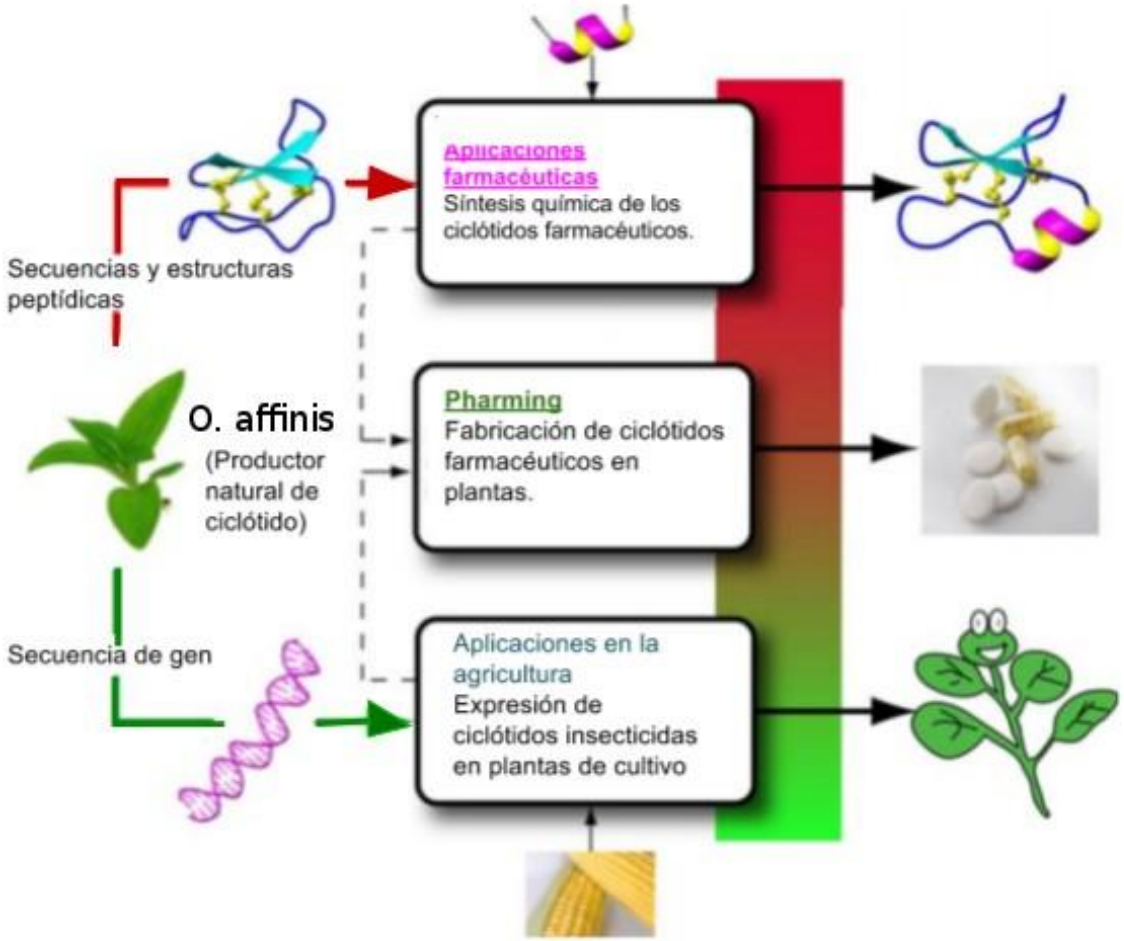


Figura 1.4: Descripción general de las aplicaciones agrícolas y del potencial farmacéutico. Fuente: Craik, David J (2010)[3].

ciclótidos o modificarlos con transformación de genes; Sin embargo para las aplicaciones agrícolas, las secuencias de los ciclótidos se expresan en ciertas plantas de cultivo para conferir resistencia a las plagas[3].

1.1.3. Ciclótidos en diferentes familias de plantas

Los ciclótidos son metabolitos secundarios existentes en ciertos grupos familiares de plantas, se encuentran en gran mayoria en la familia de las violetas (Violaceae) y la familia del café (Rubiaceae), sin embargo en los últimos años estudios afirman encontrar ciclótidos, aunque en pequeña cantidad, en las familias Fabaceae, Solanaceae, Poaceae y Apocynaceae.

1.1.3.1. La familia Violaceae

Violaceae es una familia del orden Malpighiales con 22 géneros y 930 especies en total, es reconocida en el Perú por presentar 12 géneros y 61 especies, entre hierbas, arbustos y árboles [34]. El número de ciclótidos dentro de la familia *Violaceae* excede a un estimado de 900; pero día a día se van secuenciando más ciclótidos encontrados en esta familia, los ciclótidos se pueden encontrar en mayor proporción dentro de las raíces de esta especie [35].

Estudios actuales demuestran que las especies de la familia Violaceae han desarrollado mecanismo de biosíntesis similares a la biosíntesis de los ciclótidos, tanto es el vínculo de esta familia con estas proteínas cíclicas que el estudio de solo una especie llamada Viola hederacea Labill se ha encontrado que contiene más de 50 diferentes tipos de ciclótidos [36].

1.1.3.2. La familia Rubiaceae

Rubiaceae es una de las familias de plantas más extensas que se conoce y alberga entre ellas a especies muy conocidas como es el café y la gardenia, sin embargo solo una pequeña proporción de dichas plantas contienen ciclótidos; el descubrimiento original de la Kalata B1 dentro de la región del Congo, data de uno de los primeros ciclótidos encontrados y fue extraido de O. affinis una planta de la familia Rubiaceae [37], también los ciclótidos Circulin A y B que se encuentra dentro de las plantas de la especie Chassalia parvifolia que también son Rubiaceae fueron los primeros ciclótidos reportados como agentes inhibidores de la replicación y los efectos citopáticos del VIH [38].

1.1.3.3. Fabaceae, Solanaceae, Poaceae y Apocynaceae nuevos descubrimientos

La familia fabaceae es la tercera familia más grande, comprende más de 19000 especies y representa el 27 % de la producción agrícola en todo el mundo, la principal función de estas plantas es el mejoramiento de suelos a partir del aporte de nitrógeno; los primeros ciclótidos encontrados dentro de esta familia fueron de la especie Clitoriaternatea ya que este se usaba como medicina tradicional para ayudar a las madres en el momento del parto; y se estudio ya que tiene gran similitud con la Oldenlandia affinis de la cual se han aislados muchos ciclótidos [39].

Otro tipo de familia importante donde los ciclótidos son encontrados pero de forma mínima son en la familia Solanaceae [40], Poaceae [41] y Apocynaceae [42].

1.1.4. Ciclótidos dentro de la biotecnología

Uno de los campos donde la biotecnología desempeña un papel muy importante es en encontrar terapias innovadoras de muchas enfermedades; para dicha actividad los péptidos, las proteínas y los anticuerpos monoclonales proporcionan una buena base para la investigación en ingeniería de péptidos [43]. La medicina moderna juega con las proteínas recombinantes y naturales las cuales son extraídas de plantas, su síntesis total no representa una opción económicamente rentable; una alternativa que nos da la biotecnología es el bioproceso de células vegetales hacia cultivos de suspensión en la producción de ciclótidos [44]. Otro aporte de la Biotecnología es el uso de bacterias genéticamente modificadas para la producción de ciclótidos in vivo, en un estudio realizado en el 2007 fundamentan la producción de MCoTI-II en células de E.coli modificadas genéticamente [45].

Los ciclótidos son péptidos pequeños de estructura cíclica, con seis residuos conservados de cisteína enlazado por puentes de disulfuro formando un nudo(CCK); son muy estables y aceptan cambios en su estructura, lo cual es de gran interés científico y de aplicación en la ingeniería de péptidos, una importante rama de la biotecnología [46].

1.2. Fundamentos básicos de la filogenia

En la tierra hay entre 5 y 100 millones de especies vivas de organismos, hay evidencia que sugiere que todos estos organismos están genéticamente relacionados; estas relaciones genéticas son representadas por un árbol evolutivo llamado el árbol de la vida o filogenia que es la historia del linaje del organismo. Estos árboles evolutivos o filogenias son las estructuras básicas necesarias para pensar con claridad acerca de las diferencias entre las especies [47].

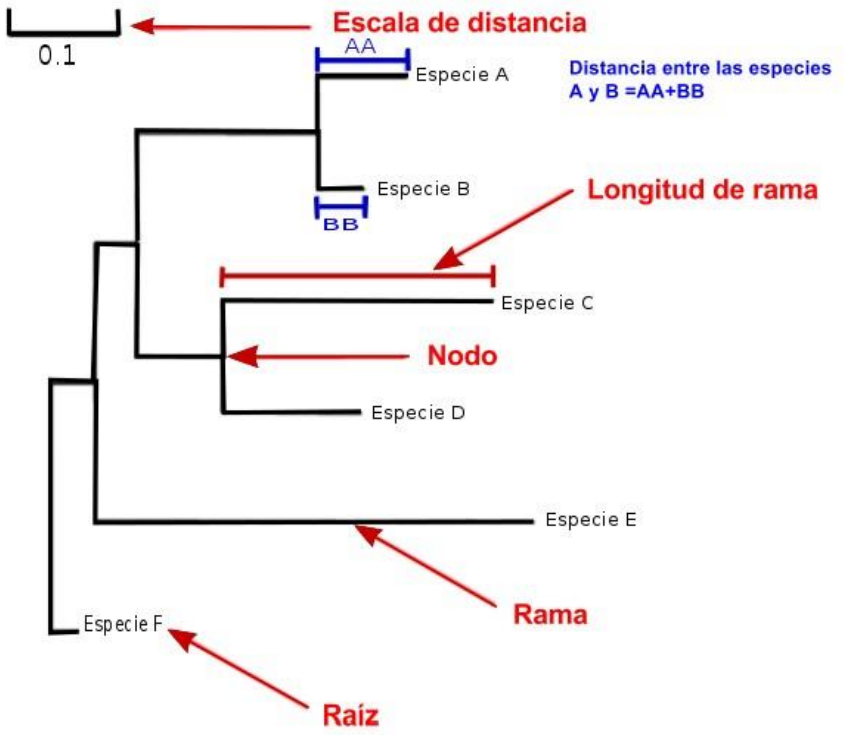


Figura 1.5: Terminología del árbol filogenético.Identificando el nodo, la rama, la topología, longitud de la rama, raíz y la escala entre distancias. Fuente: Wilkinson and Mark and McInerney (2007)[4].

1.2.1. Conceptos básicos

La finalidad de los análisis filogenéticos es estimar una filogenia que muestre la historia evolutiva del grupo taxonómico de estudio, es decir, el objetivo final es un árbol filogenético que sea reflejo del proceso de evolución donde las entidades biológicas son el resultado de "descendencia con modificación"[48 49].

La terminología de los árboles como se puede observar en la Figura 1.5 es muy simple, llamamos Nodo a una unidad taxonómica o un antepasado, la Rama define la relación existente entre los taxones según términos de ascendencia y descendencia, la Topología es el patrón de ramificación también conocido como el aspecto del árbol, la Longitud de Rama representa el número de cambios ocurridos en esa rama, la Raíz es el antepasado común de todos los taxones, finalmente la Escala entre Distancia (Distance Scale) es la escala que representa el número de diferencias entre secuencias; en donde p.0,1 significa diferencias del 10 % entre dos secuencias.

Según la existencia de rama los árboles filogenéticos se pueden clasificar en dos, enraizados y no enraizados (Figura 1.6); los árboles con raíz tienen un nodo ancestral desconocido que se puede observar, mientras que los árboles sin raíz no cuentan con dicho nodo[4]. Al construir filogenias a partir de datos moleculares es muy importante la elección del grupo externo (outgroup) ya que puede afectar de sobre manera la topología del árbol, esto se debe a que las tasas de desigualdad de evolución molecular y la topología del árbol afectan a la capacidad de los algoritmos de construcción de árboles para hallar el árbol correcto [50]

La raíz de un árbol filogenético se puede determina utilizando el método del outgroup suponiendo así que una o más especies caen fuera del grupo de interés, la rama donde se conecta el outgroup con el resto de especies formaran la raíz; también existen otros métodos para arraigar árboles filogenéticos como son el método del reloj molecular o el "midpointrooting" [51].

1.2.1.1. Información genética

Um tema trascendental dentro de la genética moderna es la variabilidad génica y el fenotipo, la información de una secuencia completa de nucleótidos en lugares específicos dentro del genoma humano revela lo diferentes que somos por individuo [52 53].

La información genética se almacena en el ADN, se transcribe en ARN y finalmente se traduce en aminoácidos; ambos ácidos nucleicos son moléculas filiformes compuestas de nucleótidos.La estructura del código genético es ahora bastante conocida cada nucleótido tiene tres componentes: un grupo químico, llamado base, un azúcar (desoxirribosa en el ADN, ribosa en el ARN) y un grupo fosfato; el código es una agrupación de tripletes formando un total de 64 trillizos los cuales más de un triplete representa cada uno de los 20 aminoácidos. en el DNA las cuatros bases son las purinas: adenina(A) y guanina (G) y las pirimidinas: timinia (T) y citocina (C), en el ARN en uracilo (U) toma el lugar de la timina [54 55].

La mayor parte de los datos dentro de las bases de datos curadas son proteicos los cuales se obtienen secuenciando y traduciendo ADN/ARN, teniendo esto en cuenta las bases de datos pueden contener errores (traducciones de marcos de lectura abiertos que no corresponden a proteínas reales, errores de secuencias de desplazamiento de trama,

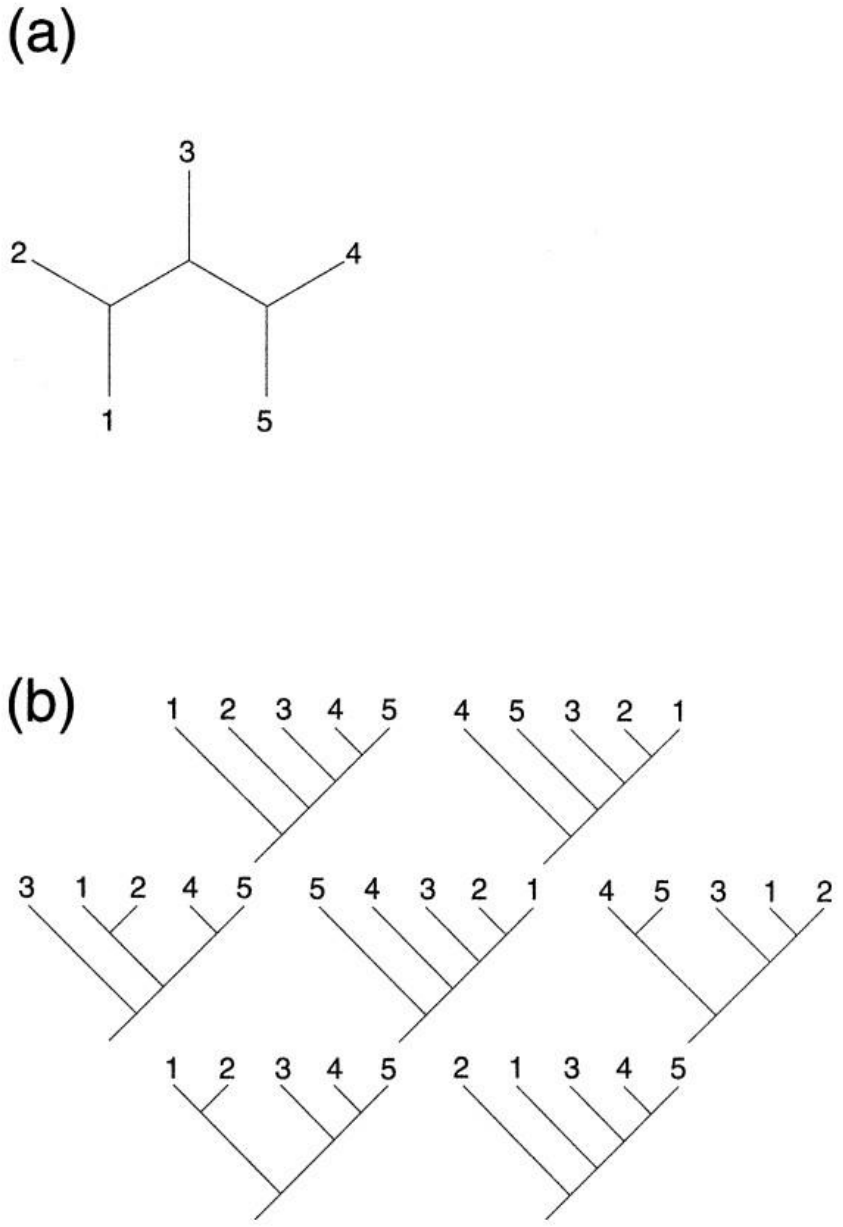


Figura 1.6: Árbol no enraizado (a) y sus siete posibles opciones de árboles enraizados (b). Fuente: Wilkinson and Mark and McInerney (2007)[4].

etc.) lo que no ocurre con los datos nucleotídicos [56]. A veces los árboles filogenéticos generados a partir de secuencias proteicas son inconsistentes y hay incertidumbre sobre la estabilidad de los datos, por lo tanto los investigadores son muy cautelosos al proponer conclusiones sobre su análisis. [57]

La deriva génica aleatoria es el nombre dado a la variación en las frecuencias genéticas que acompañan a la siguiente generación dependiendo de la formación de la generación anterior, según términos estadísticos esto corresponde a un recorrido aleatorio de las frecuencias del gen en el tiempo [58]. La importancia de la deriva génica aleatoria ha sido debatida desde que Wright propuso que la evolución es un proceso estocástico, sin embargo Fisher sustentaba que la deriva génica aleatoria es insignificante en relación con la selección natural [59]. Se conoce que la deriva génica puede tener efectos significativos en poblaciones pequeñas [60], que incluso puede dar como resultado la especiación [61], no obstante en poblaciones grandes los efectos se muestran generalmente débiles en comparación con la selección [62 63 64].

1.2.2. Cambios del material genético

Dentro del código genético una mutación en la posición del tercer codón rara vez da lugar a un cambio de aminoácidos (en un 30 %), un cambio en la primera posición en su mayoría da como resultado un cambio de aminoácidos (96 %) y un cambio en la segunda posición siempre resulta un nuevo aminoácido. Las mutaciones silenciosas o sinónimas son aquellas en las que el nucleótido cambia, pero no se observa ningún cambio fenotípico en el aminoácido, este se llama polimorfismo de un solo nucleótido o SNP; aquellos cambios de nucleótidos en los que afecta por completo al aminoácido se llama mutación no sinónima [65]. Este tipo de mutaciones proporcionan una poderosa herramienta para la compresión de los mecanismos de la evolución molecular, entre linajes la tasa de variación entre cambios sinónimos y no sinónimos pueden indicar evolución adaptativa o restricciones selectivas [66].

Las sustituciones son las mutaciones más comunes entre las bases y pueden ser agrupados en dos clases, las transiciones proceden al remplazar una pirimidina por una pirimidina (C por T) o una purina por otra purina (A por G) y las transversiones son sustituciones de una pirimidina por una purina o una purina por una pirimidina [67].

Se ha probado en estudios preliminares que mediante agentes como el etil etano sulfano-

4 o induciendo un tratamiento con PH bajo como el 6'7, podría inducir las mutaciones de transiciones o transversiones; su mecanismo de acción podría referirse a que estos agentes eliminan la guanina del ADN, resultante de este acontecimiento es que la citosina se puede incorporar a una nueva hebra de ADN que produce el original, o timina, dando lugar así a una transición también una de las dos purinas debe incorporarse ocasionalmente a la nueva hebra de ADN y provocar una Transversión [68].

Dentro del análisis filogenético las transiciones de nucleótidos son más comunes que las transversiones ya que la suposición es que las transiciones exhiben relativamente más homoplasia y por lo tanto son caracteres filogenéticos menos confiables, la homoplasia se relacionó con las tasas evolutivas y se sabe que es mayor para las transiciones, consecuentemente, las transiciones proporcionaron información filogenética sustancialmente más útil que las transversiones [69]. Según la complejidad del proceso de sustitución de nucleótidos, se han podido reconocer problemas con métodos clásicos de estima de sesgo de transición. Estos problemas se manifiestan porque hay una diferencia fundamental entre las razones de número de diferencias entre las secuencias y las proporciones de las tasas, y porque los métodos clásicos no son fácilmente generalizables [70].

otro tipo de mutación ocurrido por el cambio del marco de lectura se le llama inserción o deleción, en este caso se produce un cambio de lectura de los tripletes por la perdida o la adición de una o dos bases [71].

1.2.3. Conceptos de homología

La incorporación de caracteres genotípicos en la sistemática requieren una nueva forma en la que se aprueban las hipótesis filogenéticas y hacen ciertas inferencias, los términos de homología y analogía fueron acuñados por Oxven en 1848, pero son de hecho de origen predarwiniano, para entonces las homologías se interpretaban como la expresión de un plan básico transcendental de morfología idealista. Con Darwin, las homologías se convirtieron en uno de los argumentos más poderosos en favor de la evolución. Actualmente son considerados como estructuras heredadas de un antepasado común, la homología estática se refiere a cuando los datos están codificados en una matriz de carácter fijo antes del análisis filogenético, y la homología dinámica es donde las restricciones sobre posibles transformaciones se reducen y se infieren a posteriori como resultado del análisis

filogenético[72 73].

En el análisis filogenético la homoplasia conduce a interpretaciones erróneas, haciendo a los organismos homoplásicos parecer más cercanos evolutivamente de lo que realmente son, esta falsa pista evolutiva es debido a la evolución convergente, denominada analogía; una ascendencia con un ancestro común se le denomina homología. Distinguir entre homología y analogía es crucial para reconstruir filogenias[74].

Debido a los complicados procesos evolutivos una parte importante para entender la filogenia es la identificación de los genes ortólogos, ilustrados en la Figura 1.7; los ortólogos se definen como genes homólogos que descienden todos de un antepasado común y codifican proteínas con las mismas funciones en diferentes especies[5]. La relación homóloga entre los genes ortólogos resulta de la especiación, por otro lado los genes parálogos infieren a la homología por duplicación y pueden codificar proteínas con funciones similares pero no idénticas; por lo tanto estos genes no son útiles para determinar las relaciones filogenéticas[75].

1.2.4. Teoría neutralista de la evolución molecular

La teoría neutral afirma que la mayoría de los cambios evolutivos, a nivel molecular, son causados por la fijación aleatoria totalmente neutra; eso quiere decir que la variabilidad genética se mantiene en la especie por equilibrio entre la entrada mutacional y la extinción aleatoria [76]. Esta teoría no niega el papel de la selección natural, pero asume que sólo una pequeña fracción de los cambios de ADN en la evolución son de naturaleza adaptativa aleatoria. En otras palabras, la teoría neutral considera a los polimorfismos de proteínas y ADN como una fase transitoria de la evolución molecular y rechaza la idea de que la mayoría de estos polimorfismos son adaptativos[77].

La tasa de evolución molecular es constante en comparación con la tasa de evolución morfológica y a esta constancia se le llamo reloj molecular, y se considera como una prueba de la teoría neutralista de la evolución[78]. La forma más simple de explicar la teoría del reloj molecular es que predice la sustitución de aminoácidos en las moléculas de proteínas, es un proceso estocástico en el cual el número de sustituciones que ocurren entre proteínas homólogas puede estar relacionado con el tiempo evolutivo, en resumen nos dice que el número acumulado de sustituciones es aproximadamente proporcional al

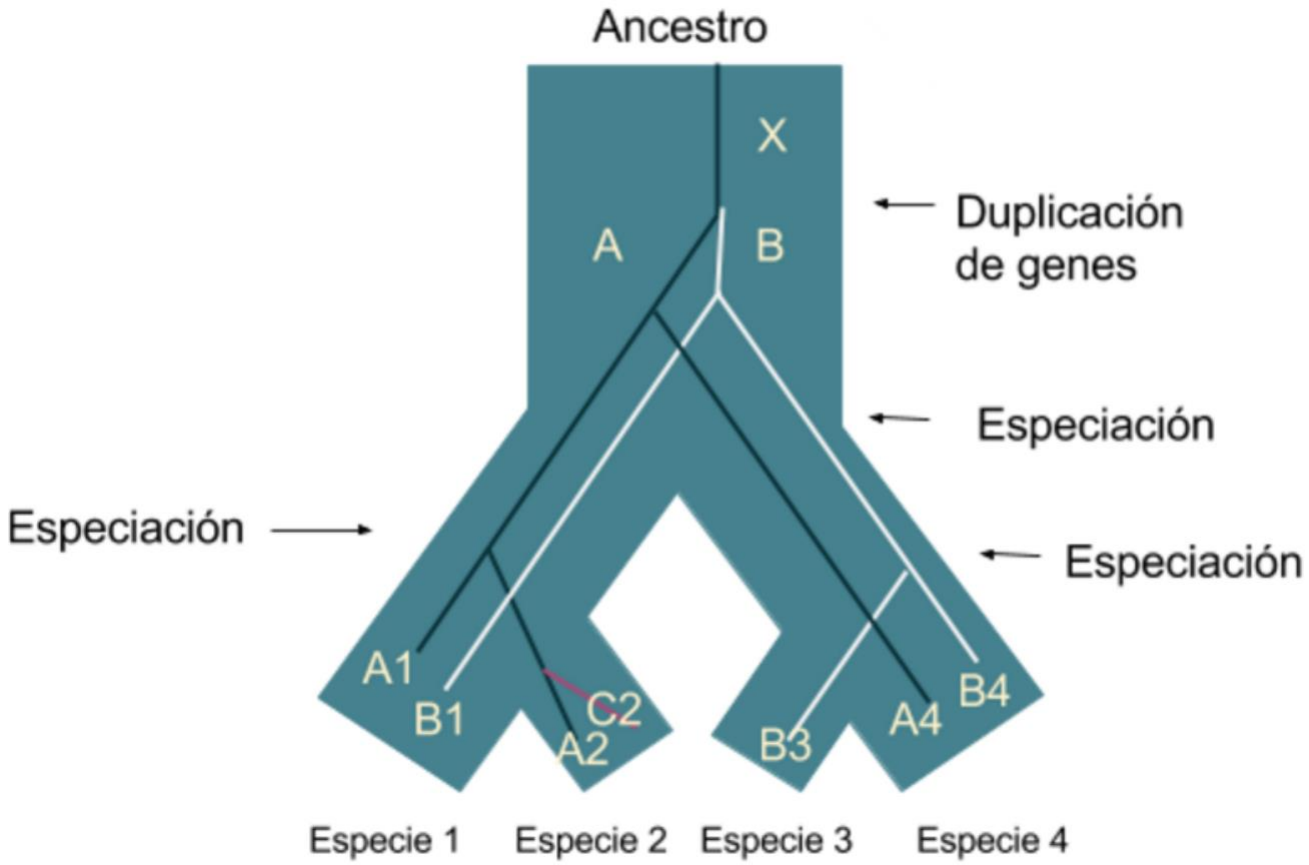


Figura 1.7: Ilustración del concepto de genes ortólogos y parálogos. Genes A y B fueron derivados por duplicación y cada uno sufrió un evento de especiación. Fuente: Bork, Peer and Dandekar (1998) [5].

tiempo [79].

1.3. Análisis filogenético

La base de todo análisis filogenético es la recopilación de secuencias proteicas o aminoacídicas de bases de datos las cuales tienen secuencias acumuladas por muchos investigadores, no es muy posible confiar en literatura impresa por lo tanto los científicos tuvieron que recurrir a las bases de datos digitalizadas. Una de las más importantes bases de datos es la NCBI entre muchas; existen también bases de datos específicas, un buen ejemplo de este es el CyBase una base de datos específica para proteínas circulares. Sin embargo la pregunta es si recopilar secuencias proteicas o aminoacídicas, la secuencia de un gen contiene toda la información necesaria para crear proteínas funcionales, y sus nucleótidos incorporan directamente las mutaciones que resultan de errores de replicación, daño por radiación, estrés oxidativo o modificación química; son razones por las que a menudo se defiende a las secuencias nucleotídicas. Sin embargo una buena defensa paras secuencias aminoacídicas es que las proteínas son las unidades fundamentales de la vida, mientras que las secuencias de ADN consisten en sólo cuatro bases A, G, C y T, las propiedades funcionales de las proteínas están determinadas por una secuencia de 20 aminoácidos posibles, lo que conduce a una resolución mucho mayor a grandes distancias evolutivas [56].

1.3.1. Alineamiento múltiple de secuencias

Al hacer un alineamiento de secuencias múltiple se garantiza obtener una matriz organizada de las secuencias en la cual los residuos de una columna dada sean homólogos, superponibles y que jueguen un papel funcional común, aunque estos criterios son equivalentes pueden divergir según su tiempo evolutivo o diferentes criterios los cuales pueden dar lugar a diferentes alineaciones [80].

El método progresivo de alineamiento más conocido es el algoritmo ClustalW, se basa en derivar primero un árbol filogenético a partir de una matriz de todas las puntuaciones de similitud de secuencias en pares, obtenido un algoritmo de alineación rápida por pares; entonces, la alineación múltiple se alcanza a partir de una serie de alineaciones de grupos de secuencias siguiendo el orden de ramificación en el árbol. El método es al parecer

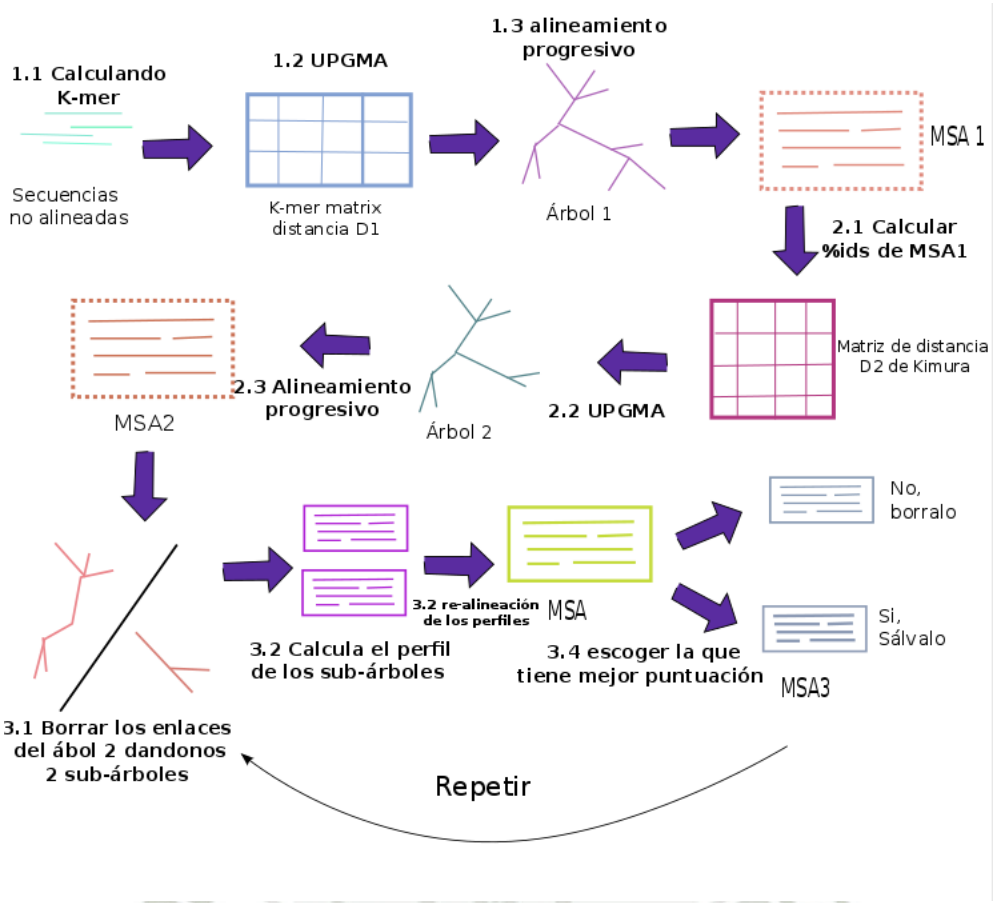


Figura 1.8: Logaritmo de MUSCLE que consta de tres etapas: primera etapa de alineación progresiva, segunda etapa de optimización del método progresivo y tercera etapa de reimplementación. Fuente: Edgar, Robert C (2004)[6].

rápido y económicamente hablando de recursos computacionales [81]. Otro método descrito para realizar el alineamiento es el T-Coffee este método se basa en la optimización del popular método progresivo y asegura ser más confiable[82].

Muscle(Figura 1.8), un método iterativo de alineación, es una de los más usados en la actualidad; los elementos del algoritmo incluyen la estimación rápida de la distancia usando el conteo del Kmer, la alineación progresiva usando una nueva función del perfil y la partición restrictiva dependiendo del árbol. Tiene las grandes ventajas de comparar la velocidad y precisión con T-Coffee, MAFFT y CLUSTALW, usarse para secuencias cuyas tasas evolutivas son muy altas y con muchos cambios moleculares y se presume de ser el método más rápido computacionalmente, alineando 5000 secuencias de una longitud de 350 aminoácidos en 7 minutos [6].

los gaps son resultantes de la alineación de secuencias cuyas longitudes son desiguales,

A

```
A ACAAT--CAGATCATCATG--ATTGT
B ACATT--CAGGTAGTCATG--AATGT
C ACATTAACAGCTAGTCATGTTAATGT
D ACAATAACAGCTCATCATGTTAATGT
E ACAAT--CAGGTCATCATG--ATTGT
F ACATT--CAGGTCGTCATG--ATTGT
```

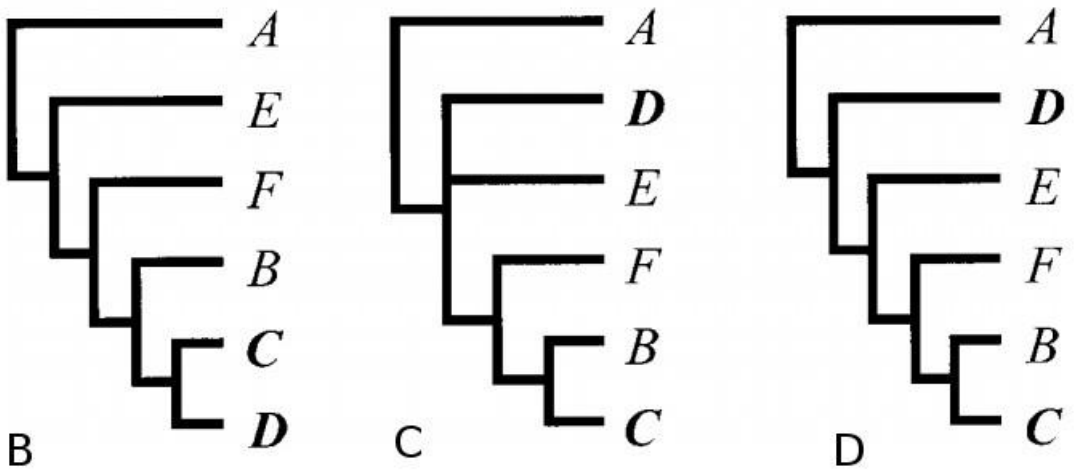


Figura 1.9: (A) la representación hipotética de secuencias; Y sus diferentes topologías (B) Parsimonia cuando los gaps son usados caracteres; (C) Parsimonia cuando los gaps son tratados como información faltante; (D) topología hecha por análisis de maximum likelihood. Fuente: Giribet, Gonzalo and Wheeler, Ward C (1999)[7].

son vistos como caracteres que se originan de eventos biológicos particulares como la mutación, los gaps contienen información histórica necesaria para el análisis filogenético o puede variar su topología como se puede observar en la Figura 1.9, el efecto de los gaps como fuente de datos filogenéticos se explora a través del análisis de sensibilidad y la congruencia de caracteres entre diferentes particiones de datos. Se proporcionan conjuntos de datos de ejemplo para mostrar que los gaps contienen información filogenética importante no recuperada por aquellos métodos que omiten los gaps en sus cálculos[7].

En el análisis de matrices basados en secuencias, se concluye que el uso de diferentes métodos de tratamiento de gaps influye considerablmente en el análisis de la hipótesis filogenéticas final; a pesar de esta influencia, un método bien justificado, aplicado uniformemente para tratar los gaps es la base para tratar todas las gaps como caracteres

separados o estados de carácter [83].

1.3.2. Modelo de evolución apropiado

Los modelos probabilísticos de evolución se emplean para las correlaciones filogenéticas; los modelos de evolución de proteínas o de nucleótidos, describen las probabilidades de cambio y por lo tanto se convierte en una herramienta fundamental para el modelamiento de evolución[84]. La selección del, mejor modelo para nucleótidos es el programa ModelTest y su optimización JModelTest2 implementa diferentes criterios estadísticos para seleccionar modelos de sustitución de nucleótidos basados en ¨Phyml¨ un algoritmo simple, rápido y preciso para estimar filogenias por máxima verosimilitud, que incluye pruebas de razón de verosimilitud jerárquica y dinámica, su metodología se muestra en la Figura 1.10, las pruebas dinámicas de razón de verosimilitud proporciona un rango de modelos de acuerdo con el criterio de información Akaike (AIC), con el criterio de información bayesiano (BIC) o a un enfoque basado en el rendimiento basado en la teoría de decisiones (DT) [85 8].

El software ProTtest es un programa java el cual se basa en el programa Phyml; utiliza la biblioteca PAL para el manejo de alineaciones de proteínas y árboles, su funcionamiento básico lo podemos encontrar en la Figura 1.11. Para los modelos sustitución de proteínas constan de una matriz de entrada de 20 x 20, diferentes matrices empíricas con las tasas relativas de sustitución de aminoácidos; entre tantos modelos elegir uno es lo más complicado, para esto esta ProtTest que consta de 112 modelos de evolución y 3 parámetros +I, +G y +F; la +I nos da la información de aminoácidos invariables, la +G asigna a cada sitio una probabilidad de pertenecer a categorías de tasas dadas y +F considera las frecuencias de aminoácidos observadas. En la selección de modelos se busca la exactitud y la sencillez, entre sus marcos de lectura esta AIC creada por Akaike que encontró una relación simple entre la probabilidad (L) y el número de parámetros (K) (Ecuación 2.1)[9]:

$$AICc = -2n \ln L + 2K \qquad (1.1)$$

Cuando el tamaño de la muestra es muy pequeño se recomienda usar AIC corregido

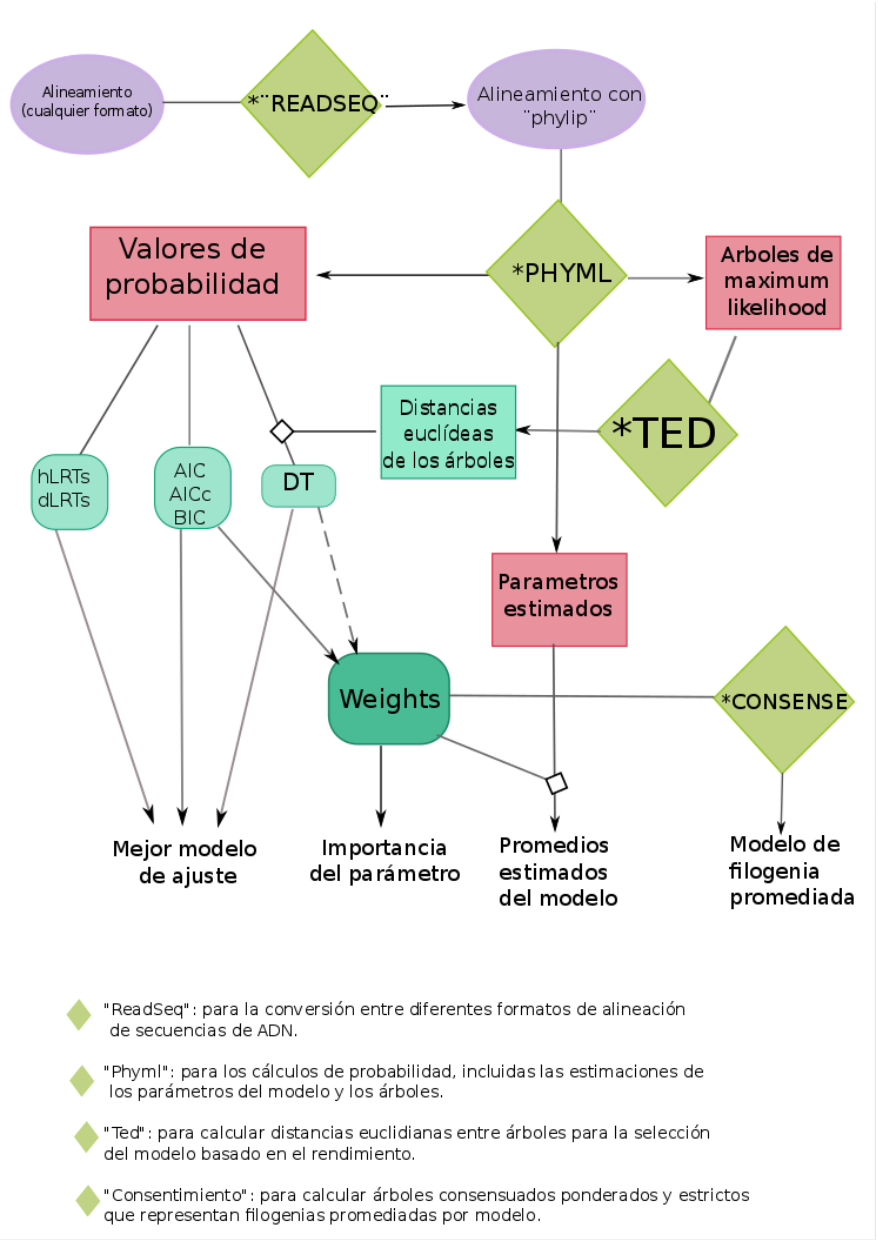


Figura 1.10: Diagrama de flujo del funcionamiento del programa JModelTest. Fuente: David Posada(2008) [8].

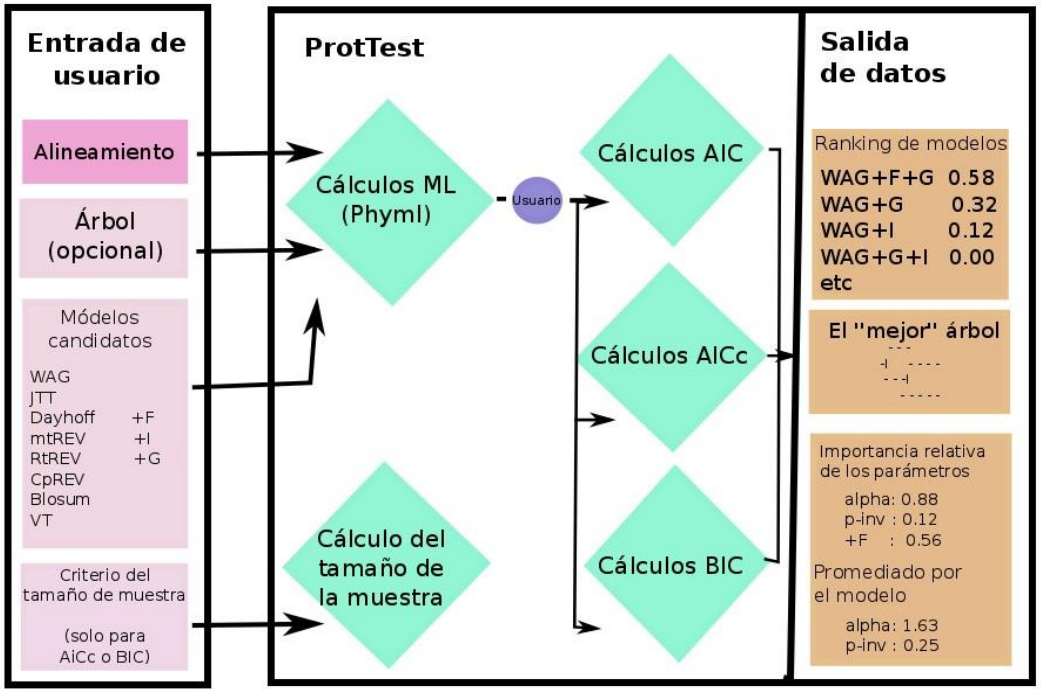


Figura 1.11: Diagrama de flujo básico del funcionamiento del programa ProtTest. Fuente: Federico Abascal, Rafael Zardoya and David Posada (2005)[9].

AICc (Ecuación 2.2)[86]:

$$AICc = \frac{AIC + (2K(K+1))}{N - K - 1} \tag{1.2}$$

Un enfoque más usado es el de criterio bayesiano BIC (Ecuación 2.3)[87].

$$BIC = -2 \ln L + K \log n \tag{1.3}$$

1.3.3. Análisis Filogenética - Métodos clásicos

La filogenia es una representación de las relaciones evolutivas entre las diferentes especies existentes, antiguamente se basaba en el estudio de caracteres morfológicos, no fue hasta principios de los años 1980's con la obtención de las secuencias de las primeras pro-

teínas y genes que nació la filogenia molecular. Existen una gran variedad dde métodos para la construcción de árboles filogenéticos:

- UPGMA
- Neighbour joining
- Máxima parsimonia
- Máxima verosimilitud
- Inferencia bayesiana

1.3.3.1. UPGMA

UPGMA (Unweighted Pair Group Method using Arithmetic averages) [88] Es un método de algoritmo de agrupación de pares no ponderados que utiliza promedios aritméticos, utilizado en la construcción de arboles filogenéticos para matrices de distancia;esta técnica esta desarrollada para construir arboles individuales, la única desventaja es que comúnmente pueden derivar más de una topología a partir de los mismos datos [89] este comportamiento depende del orden en que se introducen los datos [90] [91].

UPGMA usa un algoritmo heurístico que fue definido por los doctores sneath y Robert R. Sokal en el año 1973 [92]. y el método consiste de 4 pasos secuenciales: el primer paso consiste en determinar el número esperado de sustituciones por sitio entre los tiempos de muestreo, el siguiente paso se trata de la corrección de las distancias en pares, después se hace un cluster con el algoritmo UPGMA para corregir la matriz de distancia y se recorta las ramas posteriores [93].

1.3.3.2. Neighbour joining

Otro método para reconstruir árboles filogenéticos a partir de matrices de distancia y el algoritmo de agrupamiento es Neighbour joining, el principio de este método es encontrar pares de unidades taxonómicas operativas(UTO[=VECINO]), que van minimizando la longitud de la rama en cada etapa de agrupación de UTO; comenzando el árbol en forma de estrella, una gran característica de este método, como podemos observar en la Figura 1.12 [10].

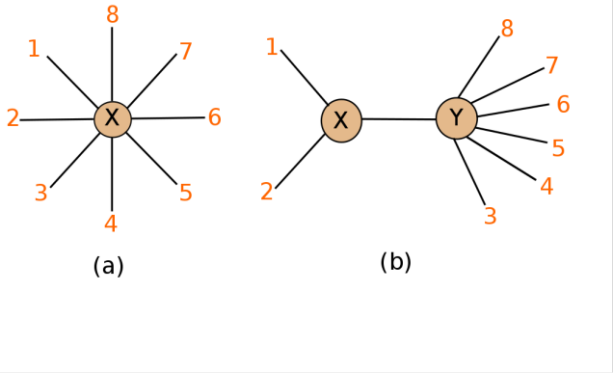


Figura 1.12: (a)Un árbol semejante a una estrella sin estructura jerárquica Y (b), un árbol en el que las UTO 1 y 2 están agrupadas. Fuente: Naruya Saitou and Masatoshi Nei (1987)[10].

Se sabe que este método usa la heurística durante el calculo de la distancia entre taxones reales y pseudo-taxones en cada paso[94] para construir un solo árbol[95] El método de Neighbour joining es útil para datos pequeños, en cual se demuestra su gran precisión y una considerable velocidad computacional;pero a medida que la cantidad de datos va creciendo, la fracción del espacio arbóreo examinada por el algoritmo de Neighbour Joining se vuelve ineficaz[96].

1.3.3.3. Minimum evolution

Los métodos de construcción de árboles filogenéticos basados en matrices de distancias muestran una correlación positiva[97]. El método de evolución mínima (ME) se basa en el supuesto de que el árbol con la suma más pequeña de longitudes de ramas es el más probable que sea el verdadero, fundamentado en el método de los mínimos cuadrados ordinarios (MCO) complementándolo con matrices de distancia; inicia construyendo una topología al estilo de Neighbour joining(NJ), luego se hace una búsqueda topológica cerca del árbol de NJ y se calcula la suma total(S) de longitudes de rama, se comparan y se elige la de menor valor de (S)[98]. Las desventajas de este método es que para un óptimo resultado deben ser muy pocas taxas por lo contrario NJ arroja mejores resultados al incrementar los taxas, también se sabe que en algunos casos dependiendo de la matriz de varianza este método es inconsistente[96].

1.3.3.4. Maxima parsimonia

Los métodos de distancia mencionados anteriormente son útiles hasta el punto en el que se pierde información, al convertir los datos originales en distancias, a diferencia de métodos probabilísticos como máxima parsimonia [99]. Sin embargo este método ha sido suplantado progresivamente, por ser inherentemente sensible al fenómeno denominado atracción de ramas largas [100]. Este fenómeno de estimaciones inconsistentes se da a partir del grado de desequilibrio de reconstruir varios estados ancestrales; en resumen el árbol no logra la estimación correcta cuando hay ramas largas en el árbol y parecen unirse, sin embargo las conclusiones de este fenómeno se basaron principalmente en estudios en los que construyeron árboles pequeños [101]. Un estudio realizado por Takezaki y Nei en el cual contaban con 5, 6 y 7, secuencias comprobaron que mientras se elevaba el número de secuencias se volvía más inconsistente el método de máxima parsimonia [102]; la inestabilidad de dicho fenómeno puede evitarse usando secuencias que evolucionan lentamente, sin embargo no es caso común en el ámbito de la investigación [103].

1.3.3.5. Maxima verosimilitud

Maximum likelihood es un método de inferencia que implica encontrar el árbol que produce la mayor probabilidad de evolucionar los datos, en este caso se quiere estimar la máxima verosimilitud y calcular la probabilidad de un conjunto de secuencias en un árbol y maximizar la probabilidad sobre todo a los árboles evolutivos [104]. Mediante estudios con datos nucleotídicos se demostró que el método de máxima verosimilitud es superior a los métodos de distancia según su eficiencia [105]. Sin embargo este método supone que la tasa de sustitución es la misma en diferentes sitios de nucleótidos, lo cual es poco realista [106].

1.3.3.6. Inferencia bayesiana

Con el desarrollo de los métodos bayesianos, es relativamente fácil explicar tanto la topología como la incertidumbre filogenética en la reconstrucción de estados ancestrales y las historias de cambio de caracteres [107]. Los avances para estimar filogenia son cada vez más novedosos, cada vez se van encontrando nuevos métodos para su estimación, la inferencia bayesiana no es un método nuevo, es uno de los métodos más antiguos de

inferencia estadística, que data del siglo XVIII. En 1968 Felsenstein habría insinuado una inferencia bayesiana de filogenia, presentando una número de ideas bayesianas, como las probabilidades posteriores de árboles y un conjunto creíble de árboles. Sin embargo, Felsenstein no fue capaz de calcular las probabilidades posteriores de los árboles en ese momento [108] [109].

Para iniciar un análisis en Una ventaja de la inferencia bayesiana en la filogenia es el aumento de la velocidad de los análisis a comparación de maximum likelihood, ademas permite realizar búsquedas más extensas de lo que era posible en un sistema basado en modelos. Esto se debe a la implementación de MCMC para estimar la distribución de probabilidad posterior, que elimina gran parte de la compleja suma e integración. La cantidad de tiempo necesario para los análisis bayesianos variará enormemente dependiendo del conjunto de datos y la metodología, pero está claro que si se quiere usar la reconstrucción filogenética basada en modelos, los análisis bayesianos pueden realizarse en un período de tiempo mucho más factible que ML [110].

La opinión sobre el valor del teorema de Bayes como metodología de inferencia estadística ha oscilado entre la aceptación y el rechazo desde su publicación en 1763; los resultados bayesianos se consideraban a veces condescendientes como un intento interesante pero equivocado de resolver un problema importante, posteriormente se encontró que las dificultades inicialmente insospechadas acompañaban a las alternativas y el interés se reavivó y en la actualidad ha vuelto a levantarse recientemente con un vigor asombroso. [111]

La inferencia filogenética con bayes se basa en una cantidad llamada probabilidad posterior, la cual se puede observar en la ecuación 2.4:

$$Pr^{h}\text{Árbol/Datos}^{i} = \frac{Pr[\text{Datos/Arbol}] \times Pr^{h}\text{Árbol}Pr^{i}}{[\text{Datos}]} \qquad (1.4)$$

Donde "/" debe ser leído como "dada" Se utiliza para combinar la probabilidad previa de una filogenia (Pr[Árbol]) con la probabilidad (Pr[Árbol de datos]) para producir una distribución de probabilidad posterior en los árboles (Pr[Árbol/Datos]). La probabilidad posterior se puede interpretar como la probabilidad de que el árbol sea correcto, eso quiere decir que el árbol con mayor probabilidad posterior puede ser elegido como la

mejor estimación de la filogenia, la probabilidad posterior es fácil de formular e implica una suma de todas los árboles y para cada árbol las posibles combinaciones de longitud de rama y valores de parametros del modelo de sustitución, la probabilidad a priori es usualmente igual en todos los árboles [112].

Los sistemáticos usan la filogenética para agrupar organismos en grupos monofiléticos, o clados, para fines taxonómicos, se optimiza este método con la cadena Markov y Monte Carlo (MCMC) para proporcionar una técnica factible desde el punto de vista computacional que satisfaga las demandas de los operadores de más taxones, manteniendo la inferencia estadística sobre una base sólida, siempre y cuando se puedan demostrar adecuadamente ciertos criterios de convergencia [113].

La cadena de Markov y Monte Carlo utiliza el algoritmo de Metropolis-Hastings para simular los procesos estocásticos, empieza con pequeños movimientos aleatorios y se evalúa la función de probabilidad, empleamos la metodología de la cadena de Markov Monte Carlo, de salto reversible (RJ), para buscar entre el gran número posible de árboles, la idea es que si repetimos este proceso muchas veces obtendremos una estimación de la probabilidad asociada a cada estado (a cada árbol) [114] [115].

En los últimos años los estadísticos han notado un gran interés en métodos como en la cadena de Markov Carlo (MCMC) para simular distribuciones multivariadas complejas y no estándar; ahora se está tomando más atención al algoritmo Metropolis-Hastings (M-H), desarrollado por Metropolis, Rosenbluth, Rosenbluth, Teller y Teller [116] y posteriormente modificado por Hastings [117]. Este algoritmo (Ecuación 2.5) es extremadamente versátil y da lugar al muestreador de Gibbs como un caso especial y usado extensamente en la física [118].

$$\pi^*(dy) = \int_{R^d} P(x, dy)\Pi(x)dx \qquad (1.5)$$

Se trata de un proceso repetido de modificar una genealogía y aceptarla o rechazarla en proporción a la proporción de su probabilidad con la probabilidad de la genealogía anterior [119].

Capítulo 2

Metodología y detalles computacionales

2.1. Detalles computacionales

2.1.1. Hardware

Workstation con procesador tipo E7 de 3.1GH, memoria RAM de 64Gb, 02 discos duros de 04 y 06TB, Acelerador de Video GTX 980 con 4GB de memoria dedicada, acelerador de video Tesla k80 con 24GB de memoria dedicada.

2.1.2. Software

ClustalW [81]:(Thompson, et al.) La función que cumple es hallar la alineación progresiva de secuencias múltiples usando el método de Matriz de distancia.

DAMBE [120] (data analysis in molecular biology and evolution):(Xuhua Xia, et al.) Es un paquete de software integrado para convertir, manipular, describir estadística y gráficamente y analizar datos de secuencias moleculares con una interfaz fácil de usar para Windows.

UGENE [121] (Okonechnikov, Konstantin, et al.): Es un software multiplataforma de código abierto con el objetivo principal de ayudar a los biólogos moleculares sin mucha experiencia en bioinformática a gestionar, analizar y visualizar sus datos. UGENE integra

herramientas de bioinformática ampliamente utilizadas dentro de una interfaz de usuario común.

JModelTest [85]:(Posada, et al.) Es un nuevo programa para la selección estadística de modelos de sustitución de nucleótidos basada en "Phyml"(Guindon y Gascuel 2003) Implementa 5 estrategias de selección diferentes, que incluyen "pruebas jerárquicas y dinámicas de la razón de verosimilitud", el çriterio de información Akaike", el çriterio de información bayesianoz un enfoque "basado en el desempeño basado en la teoría de la decisión". Este programa también calcula la importancia relativa y las estimaciones promediadas por modelo de los parámetros de sustitución, incluida una estimación promediada por modelo de la filogenia.

Mesquite [122]: (Wayne Maddison and D. R. Maddison) Es un software para la biología evolutiva, diseñado para ayudar a los biólogos a analizar datos comparativos sobre organismos. Su énfasis está en el análisis filogenético, pero algunos de sus módulos se refieren a análisis comparativos o genética de poblaciones, mientras que otros hacen un análisis multivariante no filogenético. También puede utilizarse para construir timbres que incorporen una escala de tiempo geológica, con algunos módulos opcionales.

MrBayes [123]: (J. Huelsenbeck, et al.) Es un programa para la estimación bayesiana de la filogenia. La inferencia bayesiana de la filogenia se basa en una cantidad llamada distribución de probabilidad posterior de los árboles, que es la probabilidad de que un árbol condicione las observaciones. El acondicionamiento se realiza usando el teorema de Bayes. La distribución de probabilidades posterior de los árboles es imposible de calcular analíticamente; En su lugar, MrBayes utiliza una técnica de simulación llamada cadena de Markov Monte Carlo (o MCMC) para aproximar las probabilidades posteriores de los árboles.

FigTree [124]: (Andrew Rambaut) Es un programa para ver gráficamente árboles filogenéticos. Está diseñado para mostrar archivos resumidos y anotados generados a partir de una variedad de programas, particularmente los de los archivos de salida de BEAST. El programa tiene una interfaz gráfica que permite a los usuarios modificar varios componentes del árbol, tales como posiciones de enraizamiento, etiquetas de nodos, etiquetas de punta y ejes de escala. Las cifras de árbol se pueden exportar como PDF para su publicación o para su posterior edición en otro programa de gráficos.

Tracer [125]: (Andrew Rambaut, Marc Suchard and Alexei Drummond) Es un programa para analizar los archivos generados por programas bayesianos de MCMC(es decir, para valores de parámetros continuos muestreados de la cadena). Se puede usar para analizar resultados de BEAST, MrBayes, LAMARC y posiblemente otros programas de MCMC.

BEAST [126]:(Análisis Evolutivo Bayesiano Muestreo de árboles) (Drummond, Alexei J., and Andrew Rambaut) Es un software para el análisis bayesiano de secuencias moleculares relacionadas por un árbol evolutivo. Este programa toma como entrada la extensión .xml creada por BEAUTI, proporciona una gran cantidad de modelos estocásticos populares de evolución de secuencias y se implementan modelos basados en árboles adecuados para datos de secuencia tanto dentro como entre especies.

BEAUTI [127]:(Análisis Evolutivo Bayesiano utilidades)(Rambaut, A., and A. Drummond)Es un programa de interfaz gráfica de usuario mejorado que permite el acceso a modelos avanzados para la evolución de la secuencia molecular y del rasgo fenotípico que anteriormente solo estaban disponibles para los desarrolladores.

2.2. Metodología

En el siguiente trabajo de investigación se utilizaronbas base de datos confiables NCBI(National Center for Biotechnology Information) y Cybase; se recolectó un total de 312 secuencias aminoacídicas (Figura 2.1) de ciclótidos dentro de 6 familias de plantas diferentes como son: *violaceae, rubiaceae, fabaceae, curcubitacea, poaceae y solanaceae* (ver Anexo 1) . Sin embargo, en secuencias nucleotídicas se encontraron en menor cantidad y de diferentes regiones, llevándonos a escoger aquellas secuencias de la región del mRNA del precursor del ciclótido, con el objetivo de recopilar una mayor cantidad de secuencias y tener una base de datos más amplia.

se obtuvieron secuencias de mRNA de precursores ciclotídicos en formato FASTA, de todas las familias de vegetales que expresen ciclótidos dentro de sus metabolitos secundarios reportadas hasta la actualidad; entre ellas solanaceae, poaceae, curcubitaceae, rubiace, violaceae y fabaceae. Las secuencias nucleotídicas contiene toda la información necesaria para crear proteínas funcionales, y sus nucleótidos incorporan directamente las mutaciones que resultan de errores de replicación, daño por radiación, estrés oxidativo

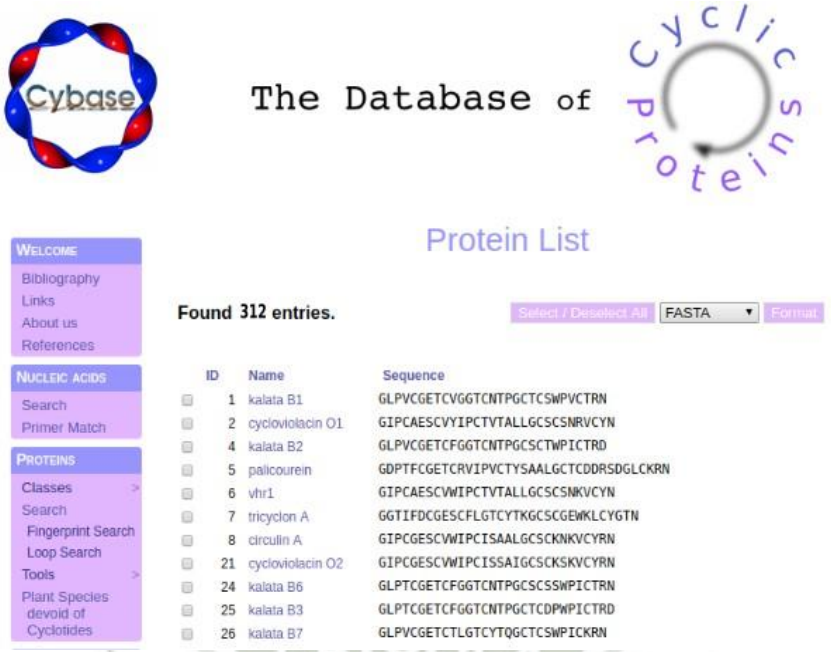


Figura 2.1: Obtención de las secuencias aminoacídicas de los ciclótidos en el servidor Cybase hasta febrero del 2017

.

o modificación química; son razones por las que a menudo se prefieren a las secuencias nucleotídicas.

El alineamiento múltiple de las secuencias se realizó con el método progresivo más popular conocido como Clustal, en este caso se usó la variante ClustalW como se puede ver en la Figura 2.2, posteriormente se curó la alineación manualmente para mejorarla y se realizó una matriz dot plot para asegurar que el outgroup es suficientemente similar al ingroup utilizando el programa UGENE.

Los datos fueron introducidos para el análisis de índice de saturación en DAMBE (Data Analysis in Molecular Biology and Evolution), se usó el test introducido por Xia para medir el índice de saturación de sustitución (iss), con el fin de evaluar si las secuencias son útiles para el análisis filogenético (Figura 2.3), a la vez se realizará el estudio de proporción de cambio entre transiciones y transversiones, con MESQUITE v3.04 se pudo hallar el punto NEXUS que contiene secuencias nucleotídicas alineadas, morfológica (estándar") de datos y sitios de restricción de datos (binarios).

Las matrices de sustitución de aminoácidos son la base para el análisis filogenético, se utilizaron para calcular las probabilidades de sustitución a lo largo de las ramas y

Figura 2.2: Secuencias alineadas mediante el programa UGENE

.

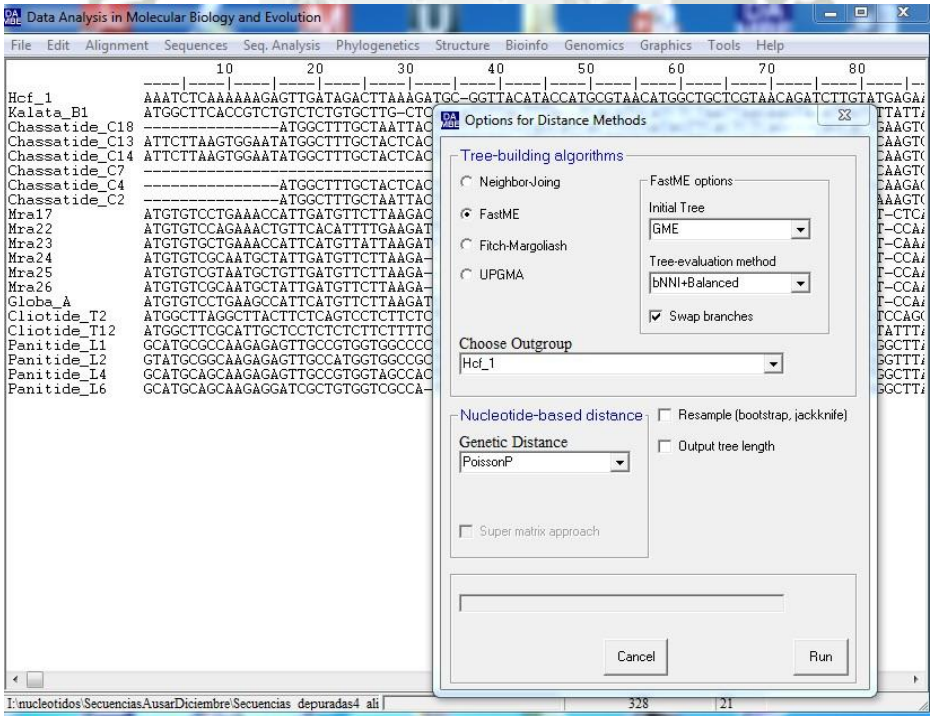


Figura 2.3: Comandos utilizados con el programa DAMBE

.

por lo tanto la probabilidad de cambio de un nucleótido a otro. Se evaluaron 64 modelos empíricos con las ocho matrices WAG, mtREV, Dayhoff, JTT, VT, Blosum62, CpREV y RtREV bajo + F, + G, + I y sus combinaciones para cada modelo, la topología del árbol se fijó en BIONJ y se eligió la estrategia de selección de modelo en BIC (Bayesian information criterion).

Se realizó la reconstrucción filogenética analizando los datos en conjunto de todas las familias seleccionadas anteriormente, se usó el análisis de inferencia Bayesiana con el programa MrBayes v3.2.3. MrBayes realiza un análisis filogenético bayesiano que combina información de diferentes datos que evolucionan bajo diferentes modelos evolutivos estocásticos. Esto permite al usuario analizar conjuntos de datos heterogéneos consistentes en diferentes tipos de datos, por ejemplo, morfológico, nucleótido y proteínico, explorar una amplia variedad de modelos estructurados mezclando partición única y parámetros compartidos. Las probabilidades posteriores de cada rama del árbol se aproximaron utilizando el método de Cadenas de Markov y Monte Carlo (MCMC) con el algoritmo Metropolis Hastings , para condicionar la lectura contando la frecuencia de árboles que presentan la misma correspondencia de taxones durante el curso del análisis. Se corrió el análisis usando entre 20 a 25 millones de generaciones, teniendo en cuenta un burnin del 25 % y 4 cadenas (una fría y tres calientes) para el diagnóstico. Se analizó las estimaciones de las divergencias temporales y los intervalos superiores e inferiores al 95 % de las densidades posteriores mayores (HPD) de estos parámetros al igual que las medias, las medias geométricas, medianas, densidades marginales y trazas fueron también estimados con el programa Tracer v 1.5. Finalmente se visualizó el árbol con el programa FigTree v1.4.0

El tiempo de divergencia para este análisis se estimó mediante el programa BEAUTI y BEAST usando aproximación del reloj molecular relajado, ya que los datos reales usualmente no se ajustan a un reloj molecular estricto, implementado en el programa BEAST v.1.8.1. Se usó el modelo de sustitución GTR con distribución gamma, para los parámetros del árbol se usó el de crecimiento exponencial y de tamaño constante de la población. Cada análisis se ejecutó a partir de un árbol al azar, de 10000000 de generaciones asegurando suficiente muestreo de parámetros. Los arboles fueron muestreados cada 100 generaciones. El primer millón de árboles fueron eliminados verificándolo con el Software Tracer para su posterior análisis.

Capítulo 3

Resultados y Discusión

Una de las partes que demanda un mayor trabajo, pero de suma importancia para el análisis filogenético, es construir una base de datos con las secuencias a analizar, razón por la que se realizó un estudio bibliográfico de diferentes bases de datos confiables para llegar a las secuencias consenso.

Se obtuvo un total de 62 secuencias nucleotídicas de la región del mRNA (ver Anexo 2) del precursor del ciclótido verificadas por el NCBI y el Cybase, esta última es una base de datos especializada en proteínas cíclicas. De la compilación final, fueron elegidas para 20 secuencias para el estudio filogenético como se muestra en la Tabla 3.1, dicha depuración se llevó a cabo mediante criterio de ruido dentro de la información filogenética. Si las secuencias son muy largas o muy cortas al punto que la información es poco relevante, estas generan ruido haciendo imposible el estudio filogenético. Las secuencias encontradas tienen un rango de 300 a 900 nucleótidos, la disimilitud en longitud de dicha secuencias hace imposible un buen alineamiento, por lo tanto al ver una gran divergencia entre secuencias se tomó la decisión de poner límites y solo usar secuencias que estén dentro del rango entre 400-500 nucleótidos intentando tener la mayor cantidad de secuencias para el estudio.

Se enraizó el árbol mediante el método del Grupo externo o Outgroup escogiendo como tal el ciclótido Hcf-1 (Tabla 3.2), con las características de ser un ciclótido hipotético derivado de una secuencia precursora asumiendo puntos de procesamiento similares a otros ciclótidos, su similitud con el grupo interno tiene un rango de 30 a 40 % siendo idóneo para este tipo de análisis como se ve en la Tabla 3.4.

Tabla 3.1: Resultado de la depuración bibliográfica de los ciclótidos a usar en el estudio filogenético; con su respectiva familia, especie, número de acceso y número de nucleótidos

Ciclótido	Familia	Especie	N° Acceso	N° Nucleotidos
kalata B1	*Rubiaceae*	*Oldenlandia affinis*	FJ211184.1	456
chassatide C18	*Rubiaceae*	*Chassalia chartacea*	JQ309971.1	433
chassatide C13	*Rubiaceae*	*Chassalia chartacea*	JQ309966.1	455
chassatide C14	*Rubiaceae*	*Chassalia chartacea*	JQ309967.1	447
chassatide C7	*Rubiaceae*	*Chassalia chartacea*	JQ309964.1	440
chassatide C4	*Rubiaceae*	*Chassalia chartacea*	JQ309963.1	468
chassatide C2	*Rubiaceae*	*Chassalia chartacea*	JQ309962.1	429
Mram 6 (Mra17)	*Violaceae*	*Melicytus ramiflorus*	EF103467.1	433
Mram 9 (Mra22)	*Violaceae*	*Melicytus ramiflorus*	EF103472.1	472
Mram 10 (Mra23)	*Violaceae*	*Melicytus ramiflorus*	EF103473.1	468
Mram 11 (Mra24)	*Violaceae*	*Melicytus ramiflorus*	EF103471.1	465
Mram 12 (Mra25)	*Violaceae*	*Melicytus ramiflorus*	EF103470.1	472
Mram 13 (Mra26)	*Violaceae*	*Melicytus ramiflorus*	EF103474.1	480
Globa A	*Violaceae*	*Gloeospermum blakeanum*	GQ438777.1	491
cliotide T2	*Fabaceae*	*Clitoria ternatea*	JF931989.1	494
cliotide T12	*Fabaceae*	*Clitoria ternatea*	JF931996.1	478
Panitide L1	*Poaceae*	*Steinchisma laxum*	KC182530.1	456
Panitide L2	*Poaceae*	*Steinchisma laxum*	KC182531.1	473
Panitide L4	*Poaceae*	*Steinchisma laxum*	KC182529.1	489
Panitide L6	*Poaceae*	*Steinchisma laxum*	KC182533.1	481

Tabla 3.2: Recopilación de la información del Grupo externo(Outgroup) a usar en el estudio filogenético

Ciclótidos	Familia	Especie	Secuencia	N° de acceso	N° de nucleótidos
Hcf-1	*Rubiacea*	*Hedyotis centranthoides*	CGAAATCCAATTCAAAAAAGAGTTGATGGGTAGACTTAAAGGAAT TCCATGCGGTGAGAGTTGCCATTACATACCATGCGTAACATCCGC GATCGGCTGCTCGTGCAGAAACAGATCTTGTATGAGAAATGAATT AACCCCTGCTGCTACATATGAAACAGACTGAAACGAGACCAAGGA ACAGAAATCTATGATGTTGCTGTTGGGATTATGTTTTCACAGTAA CCCAAATTCTGTGTGTGTTTGTGTGTGTGTTGCATTTTTTTCCTT TTTCCCTTTCCATCTTCACAAAGAGATCTTTTGATCTCCACTTGT TGTGCTTCTATTTTGTGTATCCTTTACCCTTTCGGTGTTGTTGTT CACTTGTTTGTTGCCTAAAGTGGAATAAATTTGTTATTGTTGTAA TCTTGTATCGTCTTCTATTTTCAATGTAAGTTAGTCGATTATATA AGCTCTTCTT	CB083237.1	490

Se realizó el alineamiento múltiple por matriz de análisis de las secuencias en formato FASTA mediante el algoritmo ClustalW, el método es rápido y económico hablando de recursos computacionales [81]. Existen otros métodos como MUSCLE y T-cofee los cuales también fueron probados, todos pasaron por el proceso de alineamiento mediante el programa DAMBE, seguido por el servidor G-Bloqs y finalmente curados manualmente, dejando una cantidad de gaps normalizados, los gaps constituyen una fuente valiosa de información filogenética[7] y por lo tanto se consideraron dentro del análisis filogenético. Finalmente se escogió el método ClustalW con un resultado que consta de un menor número de gaps y más sitios conservados.

Tabla 3.3: Leyenda de Ciclótidos a usar en la Tabla 3.4

Código	Ciclótido
A	Hcf_1
B	Kalata_B1
C	Chassatide_C18
D	Chassatide_C13
E	Chassatide_C14
F	Chassatide_C7
G	Chassatide_C4
H	Chassatide_C2
I	Mra17
J	Mra22
K	Mra23
L	Mra24
M	Mra25
N	Mra26
O	Globa_A
P	Cliotide_T2
Q	Cliotide_T12
R	Panitide_L1
S	Panitide_L2
T	Panitide_L4
U	Panitide_L6

Tabla 3.4: Resultado de la similitud en porcentajes comparando los 21 ciclótidos y el outgroup usando el programa UGENE

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
A	100 %	29 %	35 %	36 %	37 %	33 %	35 %	34 %	37 %	42 %	37 %	36 %	35 %	37 %	38 %	32 %	32 %	42 %	42 %	38 %	42 %
B	31 %	100 %	36 %	40 %	39 %	36 %	33 %	37 %	36 %	38 %	40 %	37 %	37 %	36 %	39 %	49 %	49 %	35 %	35 %	34 %	36 %
C	36 %	34 %	100 %	81 %	82 %	77 %	66 %	81 %	45 %	44 %	44 %	43 %	41 %	43 %	50 %	39 %	37 %	43 %	43 %	44 %	45 %
D	36 %	37 %	79 %	100 %	95 %	71 %	63 %	84 %	43 %	43 %	45 %	43 %	41 %	43 %	47 %	40 %	37 %	41 %	40 %	42 %	45 %
E	37 %	37 %	80 %	95 %	100 %	73 %	66 %	85 %	44 %	43 %	45 %	42 %	41 %	41 %	47 %	39 %	37 %	41 %	40 %	42 %	45 %
F	36 %	37 %	82 %	78 %	80 %	100 %	62 %	81 %	42 %	47 %	45 %	42 %	40 %	43 %	47 %	38 %	36 %	42 %	40 %	44 %	43 %
G	36 %	31 %	66 %	65 %	68 %	59 %	100 %	64 %	34 %	38 %	40 %	43 %	41 %	41 %	40 %	34 %	34 %	38 %	37 %	41 %	40 %
H	36 %	36 %	83 %	89 %	90 %	78 %	65 %	100 %	44 %	46 %	47 %	42 %	41 %	42 %	50 %	38 %	37 %	40 %	40 %	42 %	45 %
I	39 %	36 %	46 %	46 %	46 %	41 %	35 %	44 %	100 %	71 %	68 %	57 %	55 %	56 %	65 %	36 %	38 %	39 %	37 %	40 %	39 %
J	42 %	35 %	43 %	44 %	44 %	43 %	38 %	44 %	68 %	100 %	69 %	56 %	56 %	55 %	66 %	32 %	33 %	38 %	39 %	40 %	37 %
K	36 %	36 %	42 %	44 %	44 %	40 %	38 %	44 %	63 %	67 %	100 %	56 %	56 %	56 %	77 %	36 %	36 %	37 %	36 %	36 %	38 %
L	36 %	34 %	42 %	43 %	42 %	39 %	42 %	40 %	54 %	56 %	58 %	100 %	93 %	94 %	63 %	34 %	32 %	40 %	40 %	44 %	42 %
M	35 %	35 %	40 %	41 %	41 %	37 %	40 %	39 %	52 %	56 %	57 %	93 %	100 %	92 %	61 %	32 %	31 %	38 %	39 %	43 %	41 %
N	37 %	34 %	42 %	43 %	42 %	39 %	40 %	40 %	53 %	55 %	57 %	94 %	93 %	100 %	62 %	33 %	30 %	39 %	39 %	43 %	43 %
O	38 %	36 %	48 %	47 %	47 %	43 %	39 %	47 %	62 %	65 %	78 %	62 %	61 %	61 %	100 %	36 %	35 %	39 %	39 %	42 %	40 %
P	33 %	46 %	38 %	40 %	40 %	36 %	34 %	37 %	35 %	32 %	37 %	34 %	33 %	33 %	37 %	100 %	72 %	35 %	36 %	33 %	36 %
Q	32 %	46 %	36 %	38 %	37 %	34 %	34 %	35 %	36 %	33 %	37 %	32 %	31 %	31 %	36 %	72 %	100 %	35 %	36 %	32 %	35 %
R	44 %	34 %	44 %	43 %	43 %	41 %	39 %	40 %	39 %	40 %	40 %	42 %	40 %	41 %	42 %	36 %	37 %	100 %	85 %	74 %	85 %
S	44 %	34 %	44 %	42 %	41 %	38 %	38 %	39 %	36 %	41 %	38 %	41 %	40 %	41 %	41 %	37 %	37 %	85 %	100 %	71 %	78 %
T	39 %	32 %	44 %	43 %	43 %	41 %	40 %	41 %	39 %	40 %	38 %	44 %	44 %	43 %	43 %	34 %	33 %	72 %	69 %	100 %	73 %
U	43 %	34 %	45 %	47 %	46 %	41 %	40 %	44 %	38 %	38 %	40 %	43 %	42 %	44 %	42 %	37 %	36 %	83 %	77 %	73 %	100 %

Tabla 3.5: Resultados obtenidos mediante el programa JModelTest para la elección del modelo de sustitución nucleotídica para cada tipo de criterios

Parametros	AIC	BIC	DT
Tamaño de muestra	328	328	328
Modelo	TIM3+G	TIM3+G	TIM3+G
Partición	012032	012032	012032
-Lnl	5031.0753	5031.0753	5031.0753
k	47	47	47
FreqA	0.2359	0.2359	0.2359
FreqC	0.1906	0.1906	0.1906
FreqG	0.2133	0.2133	02133
FreqT	0.3603	0.3603	0.3603
R(a) [AC]	2.5565	2.5565	2.5565
R(b) [AG]	3.5000	3.5000	3.5000
R(c) [AT]	1.0000	1.0000	1.0000
R(d) [CG]	2.5565	2.5565	2.5565
R(e) [CT]	2.4085	2.4085	2.4085
R(f) [GT]	1.0000	1.0000	1.0000
Gamma shape	4.8800	4.8800	4.8800

Las matrices de sustitución de nucleótidos son la base para el análisis filogenético y se utilizan para calcular las probabilidades de sustitución a lo largo de las ramas y por lo tanto la probabilidad de cambio de un nucleótido a otro, jModelTest dio el resultado de pruebas dinámicas de razón de verosimilitud que proporciona un rango de modelos de acuerdo con el criterio de información Akaike (AIC), con el criterio de información bayesiano (BIC) y un enfoque del rendimiento basado en la teoría de decisiones (DT), como se muestra en la Tabla 3.4

Los resultados indican que el modelo TIM3+G (Tabla 3.4), es el que más se adecua para hacer el análisis conjunto de todas las familias, también se tuvieron en cuenta los parámetros G (gamma: heterogeneidad evolutiva dentro de las secuencias), I (proporción de zonas invariables), F (frecuencia de aminoácidos observados); los valores de estos parámetros son estimados en función de su valor de probabilidad a posteriori, como se puede observar en la Tabla 3.4 se toma en cuenta el criterio de información bayesiano (BIC), el cual se considera una buena aproximación de los métodos bayesianos, según Posada et al.[9] de acuerdo a este criterio los mejores modelos a escoger son los que tienes el -lnl menor o el BICw mayor. Sin embargo, al querer usar métodos Bayesianos el modelo TIM3 no se encuentra en la relación de modelos que se puedan usar, por lo tanto se usará

Tabla 3.6: Resultado de los primeros 5 modelos de sustitución nucleotídica para el riterio BIC

Model	-lnL	K	BIC	Delta	Weight	cumWeight
TIM3+G	5031.07530	47	10334.422240	0.000000	0.765176	0.765176
TPM3uf+G	5035.45821	46	10337.3 95046	2.972806	0.173071	0.938247
TIM3+I+G	5031.11440	48	10340.293453	5.871214	0.040630	0.978877
GTR+G	5029.65893	49	10343.175527	8.753287	0.009616	0.988493
TOM3uf+I+G	5035.49777	47	10343.267180	8.499940	0.009186	0.997679

el modelo GTR+G que cumple en cuarta posición del criterio bayesiano con la escala obtenido en JModelTest (Tabla 3.6).

El análisis de saturación realizado muestra la relación entre las transversiones / transiciones la distancia genética (F84), en donde se muestra que existe un aumento gradual de las tasas de sustitución nucleotídica con respecto a la distancia genética entre los ciclótidos. Esta relación no linear formado entre las variables, con una breve perturbación al inicio del gráfico, indica que existe cierta saturación propia de la naturaleza de los ciclótidos (Figura 3.1).

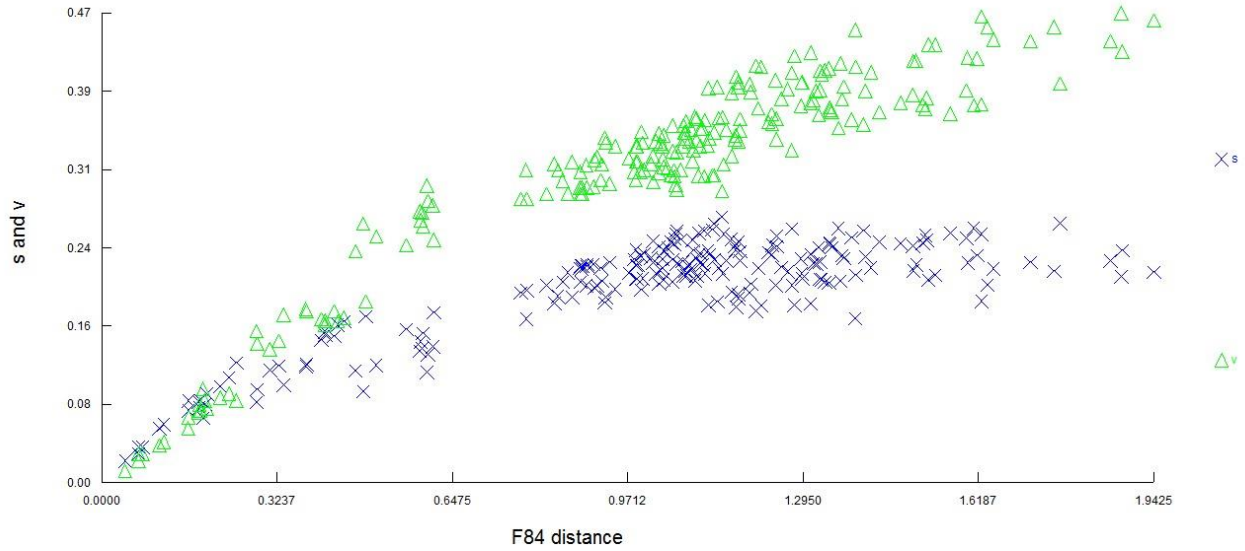


Figura 3.1: Prueba de saturación. Los triángulos en verde muestran las transversiones y las cruces en azul las transiciones. Eje de las X muestra la distancia genética corregida F84; el eje de las Y muestran el número de substituciones nucleotídicas.

La topología derivada mediante inferencia bayesiana fue ejecutada a través del programa MRbayes v3.1.2.[107]. Todos los parámetros a priori fueron introducidos en el programa,

entre ellos el modelo de sustitución nucleotídica, el cual fue seleccionado por el criterio de información bayesiana (GTR+G), se realizaron dos corridas con el fin de buscar las cadenas de Monte Carlo (MCMC).

Se evaluó el punto de convergencia de las dos corridas en conjunto mediante el programa Tracer v 1.5[125], eliminando el principio de muestreo del 25 %, conocido como burn-in, ya que los valores de verosimilitud son cambiantes y van en aumento. Se analizó el tamaño de muestra efectiva (ESS) dándonos como resultado todos los parámetros mayores a 200 (Tabla 3.7), lo que quiere decir que se ha alcanzado un adecuado número de muestras independientes; un ESS menor a 100 indicaría que muchas de las muestras están correlacionadas y por lo tanto puede no representar bien la distribución posterior. Entre los parámetros más importantes a estudiar están: LnL, que significa logaritmo natural de verosimilitud; LnPr, que significa logaritmo natural de probabilidad y TL que representa la suma de todas las longitudes de rama.

Los parámetros estadísticos como las medias, las medias geométricas, varianza, error estándar de la media y moda, para los parámetros LnL y LnPr se encuentran en las Tablas 3.8 y 3.9 correspondientemente. La incertidumbre estadística de los datos se refleja en el valor de la densidad de probabilidad más alta (HPD) al 95 % cuyos intervalos superiores e inferiores se muestran como el rango de confianza bayesiano. El número de estados en la cadena MCMC que dos muestras necesitan entre sí para que no estén correlacionadas se ve a través del tiempo de auto correlación (ACT).

El histograma mostrado en la Figura 3.2 muestra el parámetro LnL a través del tiempo, evidenciando que las cadenas han logrado estacionalidad y convergencia, la curva ha alcanzado una distribución normal o en forma de campana de Gauss y no se nota perturbación alguna, reflejando así un óptimo resultado. De la misma forma se observa para el parámetro LnPr el cual nos muestra la misma distribución natural en la curva (Figura 3.3) confirmándonos la convergencia de las cadenas.

El siguiente parámetro a analizar son las tazas de sustitución entre nucleótidos (A-C, A-G, A-T, C-G, C-T, G-T); con la densidad marginal, aquí podemos ver y comparar la densidad marginal de probabilidad posterior para todas las transiciones a la vez, se observó que dichas transiciones tienen forma de campana invadiendo ligeramente las gráficas entre sí, con lo cual se asume saturación entre las transiciones, esta densidad fue

Tabla 3.7: Parámetros estimados con el programa Tracer con sus respectivas medias y tamaño de muestra efectiva (ESS).

Parámetros	Media	ESS
LnL	-5056.376	69757
LnPr	-20.821	78397
TL	6.972	75905
r(A<->C)	0.183	53538
r(A<->G)	0.272	47775
r(A<->T)	8.566E-2	61304
r(C<->G)	0.206	51252
r(C<->T)	0.187	54236
r(G<->T)	6.653E-2	60356
pi(A)	0.235	61011
pi(C)	0.192	61006
pi(G)	0.213	54808
pi(T)	0.361	51664
alpha	4.186	78769

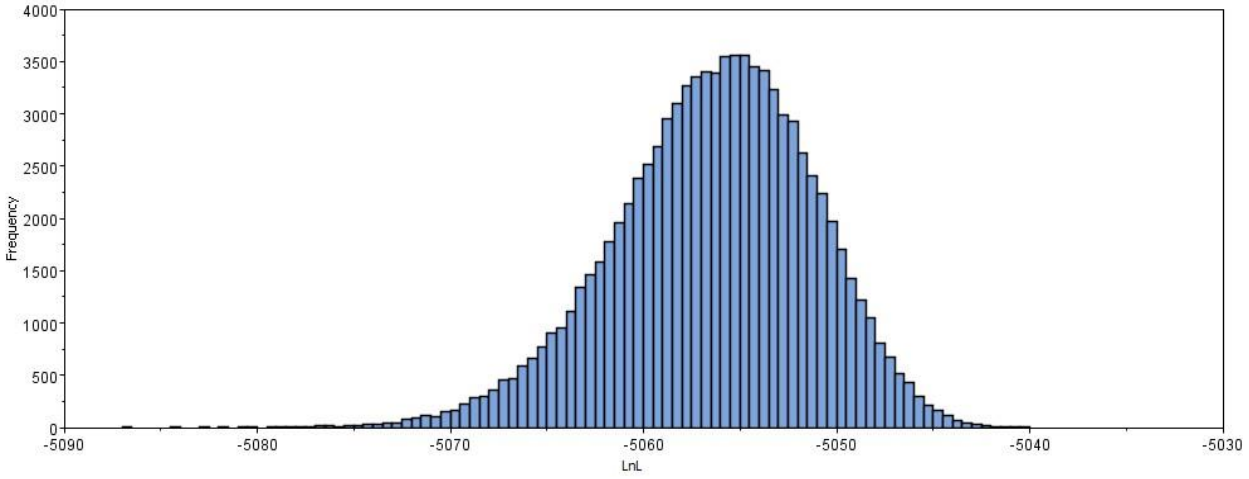


Figura 3.2: Estimación a porsteriori del LnL (logaritmo natural de verosimilitud) realizada por las dos corridas en conjunto mostrando una distribución normal, que significa stacionariedad y convergencia.

Tabla 3.8: Resumen estadístico para el parametro LnL.

Parametro estadístico	Valor
Media	-5056.3757
Error estándar de la media	0.0194
Desviación estándar	5.128
Varianza	26.3037
Mediana	-5056.046
Moda	n/a
Media geométrica	n/a
95 % intérvalo HDP	-5066.528, -5046.612
Autocorrelación de tiempo (ACT)	645.1118
Tamaño de muestra efectiva (ESS)	69756.8972

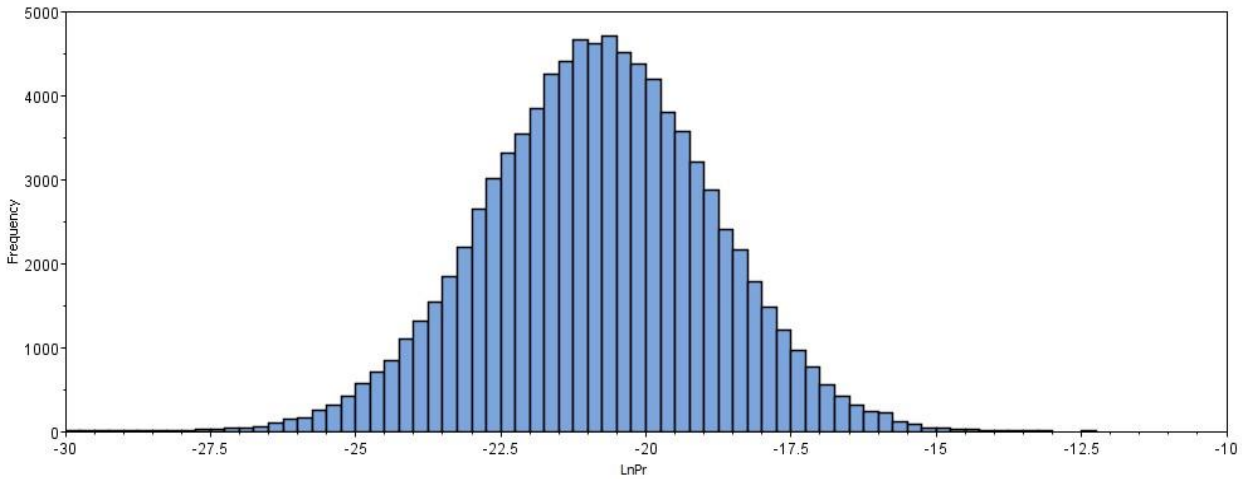


Figura 3.3: Estimación a porsteriori del LnPr (logaritmo natural de probabilidad) realizada por las dos corridas en conjunto mostrando una distribución normal, que significa stacionariedad y convergencia.

Tabla 3.9: Resumen estadístico para el parametro LnPr.

Parametro estadístico	Valor
Media	-20.8209
Error estándar de la media	6.9624E-3
Desviación estándar	1..9494
Varianza	3.8003
Mediana	-20.8029
Mode	n/a
Media geométrica	n/a
95 % intérvalo HDP	-24.6729.-17.0205
Autocorrelación de tiempo (ACT)	574.0139
Tamaño de muestra efectiva (ESS)	78397.0512

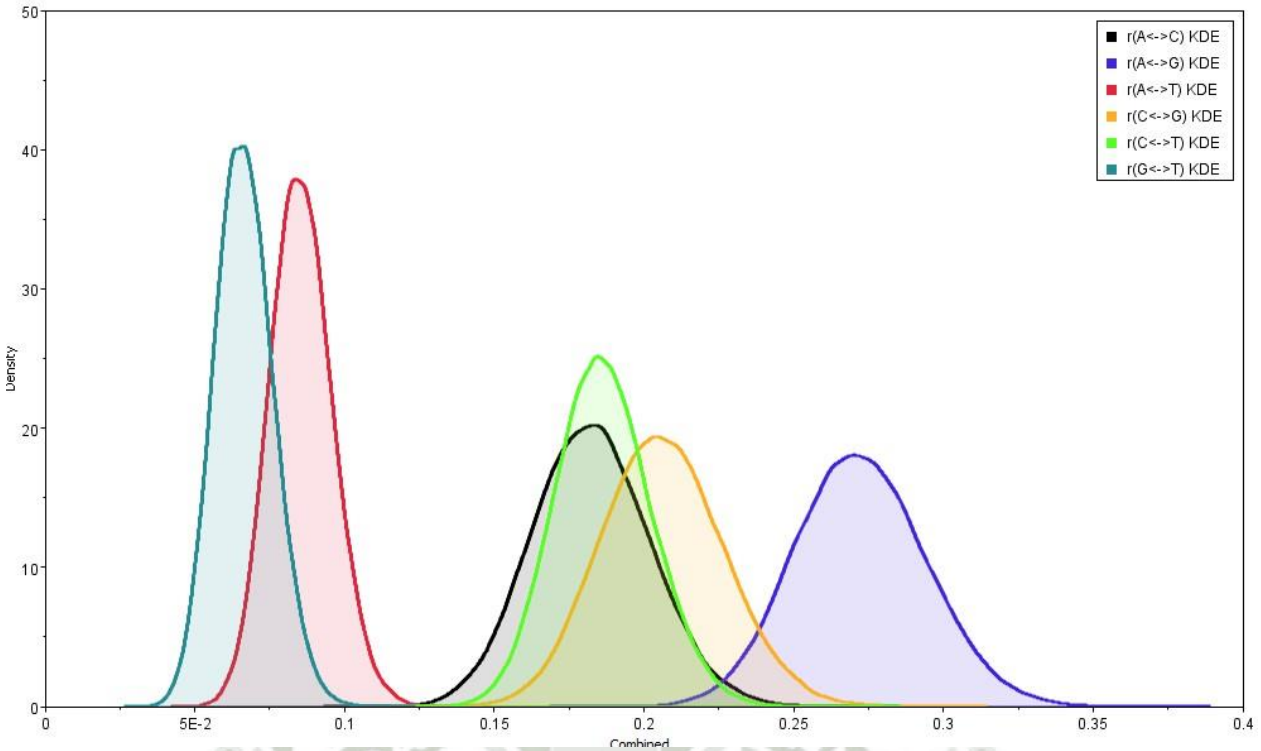


Figura 3.4: Gráfica de la probabilidad de distribución marginal de la transición de los nucleótidos mediante en parámetro KDE (kernel density estimate).

analizada con el parámetro KDE (kernel density estimate), representa un método común no paramétrico para suavizar histogramas en estimaciones de la función de densidad de probabilidad (Figura 3.4), de la misma forma fue analizado a través del histograma, en el cual se puede visualizar el poco ruido ocasionado por las cadenas (Figura 3.5).

Se analizó el gráfico conocido como trace para las dos corridas por separado y en conjunto, ambas corridas con 25000000 generaciones realizando un burn-in del 10 % para cada una (Tabla 3.10), es importante excluir las primeras muestras porque las cadenas no han alcanzado la estacionariedad en este punto [114]. En la primera corrida las estimaciones para la probabilidad posterior han subido y bajado entre -5066.528 y -5046.589 mostrando una fluctuación de gran tamaño al final del análisis (Figura 3.6), mientras que en la segunda corrida las estimaciones se encuentran entre -5066.601 y -5046.728 mostrando una mayor uniformidad (Figura 3.7). Las líneas deben desplazarse dentro de dicho rango con pequeñas variaciones en cada lado, pareciéndose a lo que los autores llaman hairy caterpillar (oruga peluda) [125], el cual es indicador que se ha ejecutado las cadena el tiempo suficiente para la convergencia.

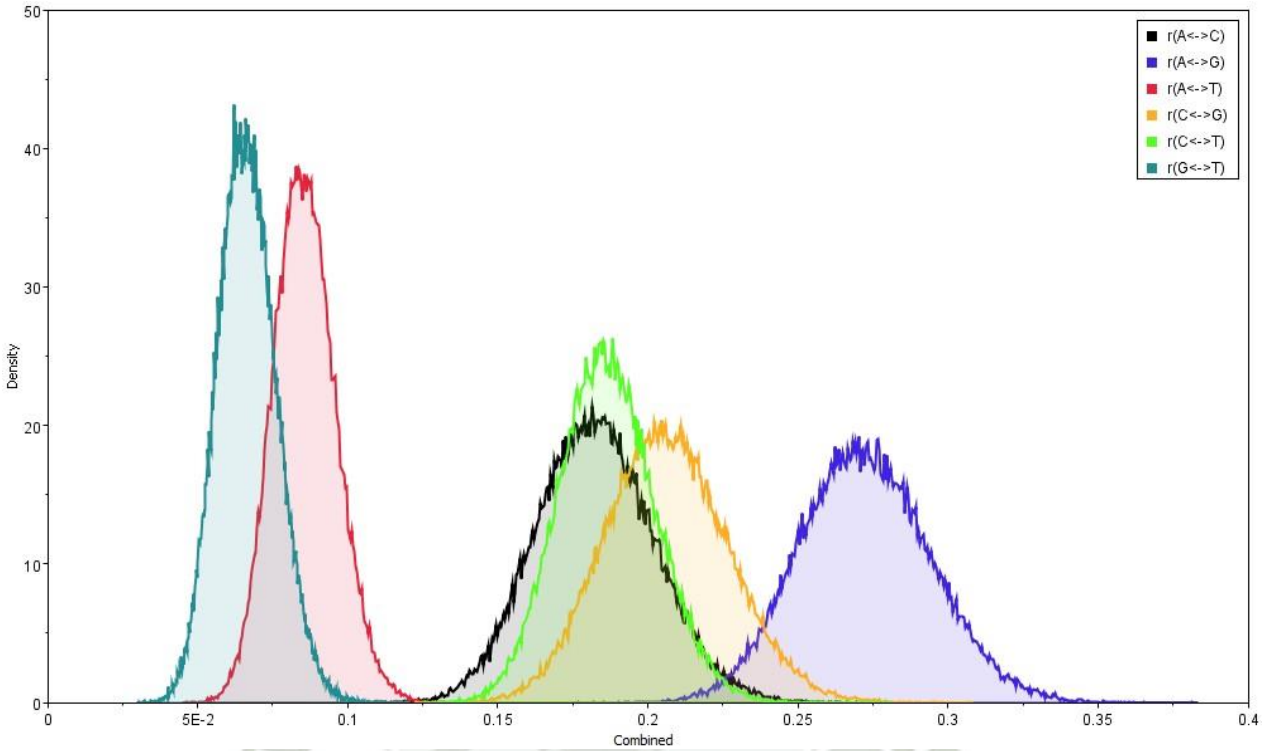


Figura 3.5: Gráfica de la probabilidad de distribución marginal de la transición de los nucleótidos mediante un histograma, para observar el ruido del análisis.

Tabla 3.10: Tabla sobre los saltos y burn-in de las cadenas en separado y en conjunto respectivamente.

Archivo	Generaciones	Burn-in
Primera Cadena	25000000	2500000
Segunda Cadena	25000000	2500000
Combinadoa	45001000	-

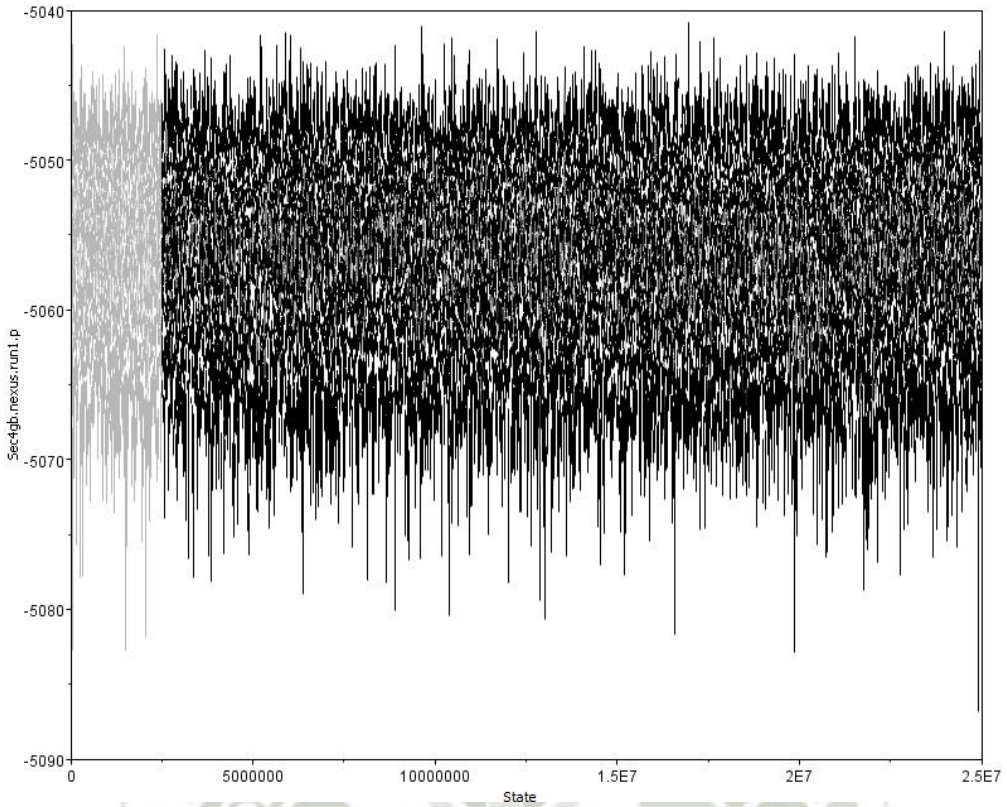


Figura 3.6: Gráfico tipo tracer comparando las cadenas MCMC de la primera corrida en el eje x contra los valores de verosimilitud logarítmica (LnL) en el eje y.

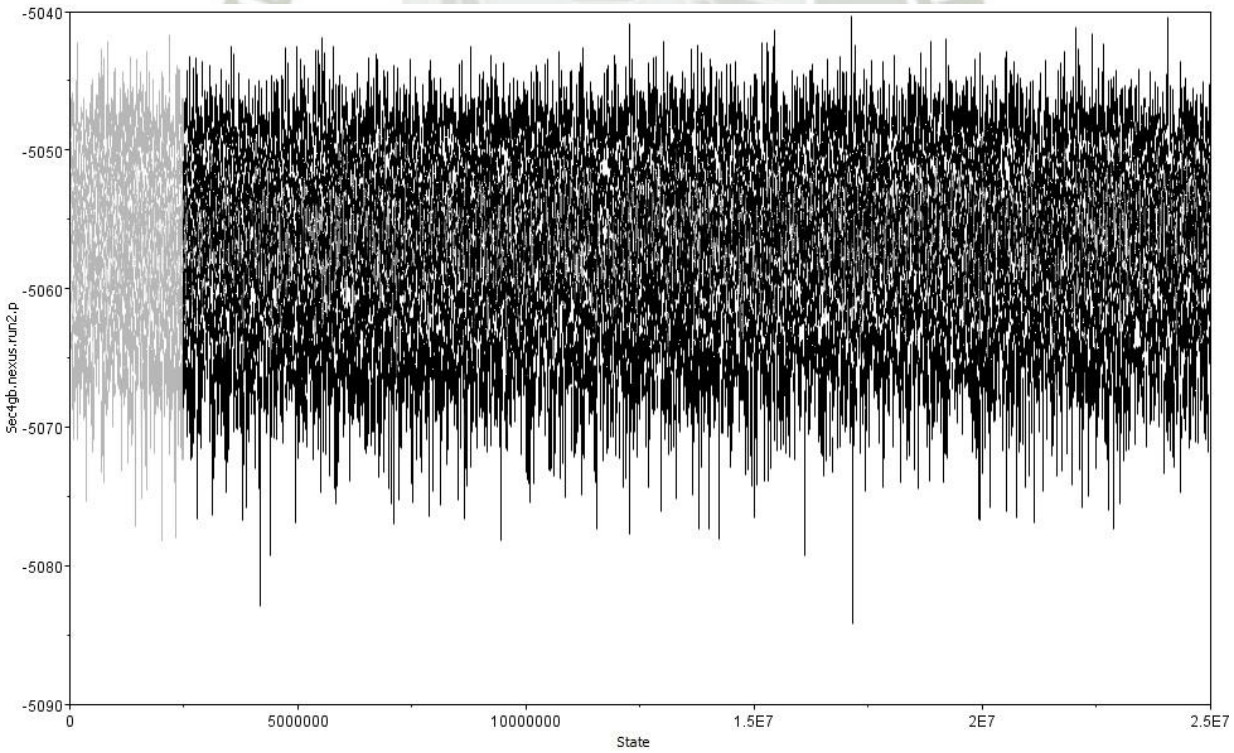


Figura 3.7: Gráfico tipo tracer comparando las cadenas MCMC de la segunda corrida en el eje x contra los valores de verosimilitud logarítmica (LnL) en el eje y.

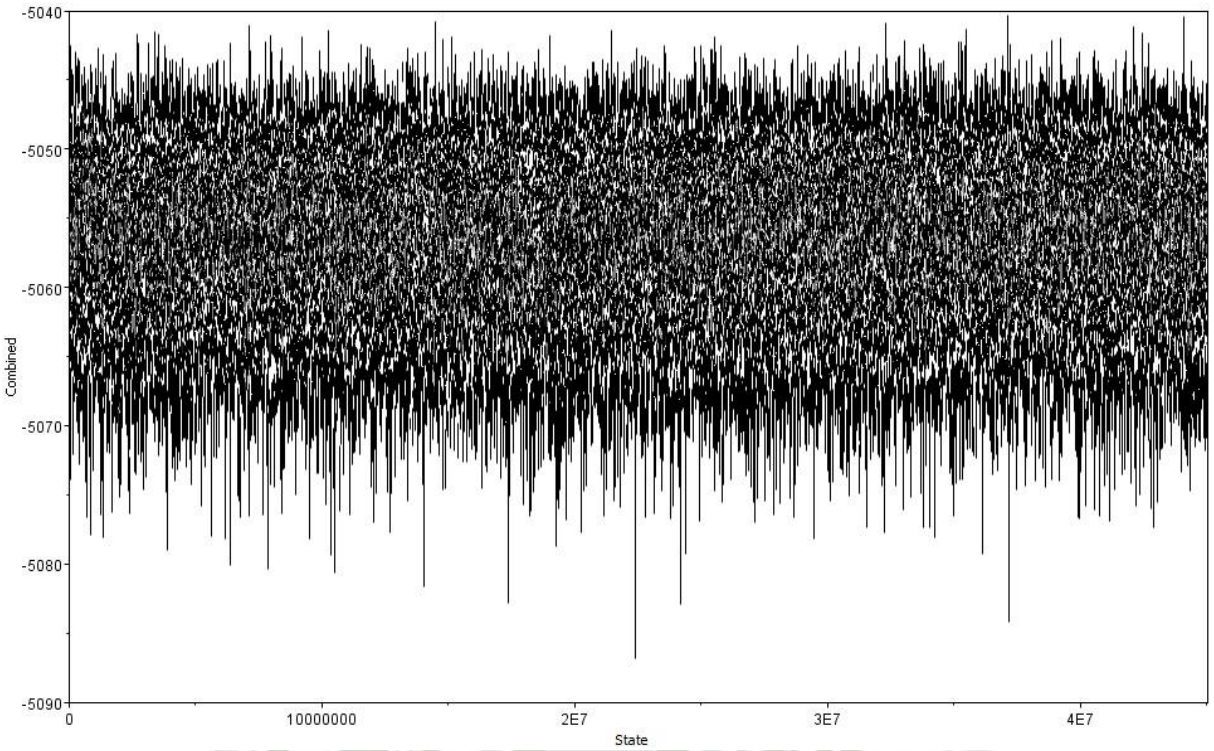


Figura 3.8: Gráfico tipo tracer comparando las cadenas MCMC de las corridas en conjunto en el eje x contra los valores de verosimilitud logarítmica (LnL) en el eje y.

En la Figura 3.8 se muestra el análisis en conjunto de las dos cadenas combinadas, el gráfico trace en forma de hairy caterpillar (oruga peluda) cuyas fluctuaciones han subido y bajado entre -5066.528 y -5046.612, evidencia que es buena señal de convergencia, lo que quiere decir que hay poca auto-correlación entre las muestras. No hay tendencias que sugieran que la MCMC aún no ha convergido, no requiere agregar más cadenas.

Se construyó un árbol con inferencia bayesiana el cual muestra una estructura compuesta por cuatro clados, cada uno con familias bien diferenciadas y con valores de probabilidad a posteriori lo suficientemente grande para soportar los clados. Los ciclótidos parecen tener altas similitudes en su secuencia y en su identidad estructural; este alto grado de homología sugiere la conservación durante la evolución, posiblemente debido a un papel importante en la defensa de las plantas contra las plagas y los patógenos [11].

El valor de credibilidad generado por el análisis se indica a través de la probabilidad posterior (PP), en la Figura 3.9 se observa dentro del rango 63,4316 % a 100 % en colores rojo y azul respectivamente, los valores más bajos se muestran dentro del Clado C el cual es una familia homóloga de *rubiaceae*, entre Chassatide c18 y c7, esto se puede deber a que

Chassatide c7 se muestra como un ciclótido lineal. En el Clado A se observa la familia de *poaceae* bien conservada, mientras que en el Clado B se identifican dos familias, Cliotide T2 y Cliotide T12 que pertenecen a la familia *fabaceae* y Kalata B1 que está dentro de la familia *rubiaceae*, ambas familias sugieren tener un ancestro en común validado por el PP con un 100 %, en estudios realizados sobre los ciclótidos de la familia *fabaceae*, muestran que sus precursores son quimeras, lo que quiere decir que tienen capacidad para producir proteínas que combinan parte de las funciones de dos proteínas distintas, sus ancestros son *rubiaceae y violaceae*, con esto sugiere que la aparición de ciclótidos en las *fabaceae* podría ser el resultado de una transferencia horizontal de genes entre los genomas nucleares de las plantas o una evolución convergente, proporcionando una nueva comprensión sobre el mecanismo biosintético de ciclótidos en las *fabaceae* y su evolución en las plantas [128].

En el Clado D se muestra a la familia *violaceae* bien conservada, mostrando un valor bajo en la divergencia entre Mra17 y Mra22 ambos son de la misma familia, pertenecen al mismo género de plantas y según su porcentaje de similitud de secuencia cuenta con un 71 %, la credibilidad baja puede deberse a que son estructuralmente diferentes debido a la función que presenta, ya que una única mutación puede conducir a la modificación de la función, explicando así como algunas proteínas están relacionadas filogenéticamente porque comparten una función común aún no probada [129]. Como se puede observar en los Clados B, C y D los tres clados contienen ciclótidos de las familias *violaceae* y *Rubiaceae* soportados con un índice de credibilidad al 89,3909 %. *rubiacea* y *violacea* son familias de plantas lejanamente relacionadas, sin embargo hay estudios que explican su cercanía filogenética, ya que los péptidos de la subfamilia brazalete y Mobius no se agrupan según su familia botánica de origen y es por qué podría haber existido una proteína ancestral común antes de la separación de las familias Rubiacea y Violacea[11].

En la Figura 3.10 nos muestra los valores superiores e inferiores dentro de la densidad de probabilidad más alta al 95 % de los ancestros de cada clado, asumiendo la edad en millones de años; este rango se muestra por las barras en azul dibujadas dentro del árbol. Entonces, [0.396,0.95] en el nodo entre Kalata B1 y la familia Cliotide significa que el 95 % de HPD para ese nodo tiene entre 0.396 años de edad y 0.95 (Mya), es el rango más grande que se puede observar dentro del árbol construido asumiendo una evolución rápida de los demás ancestro en los cicótidos.

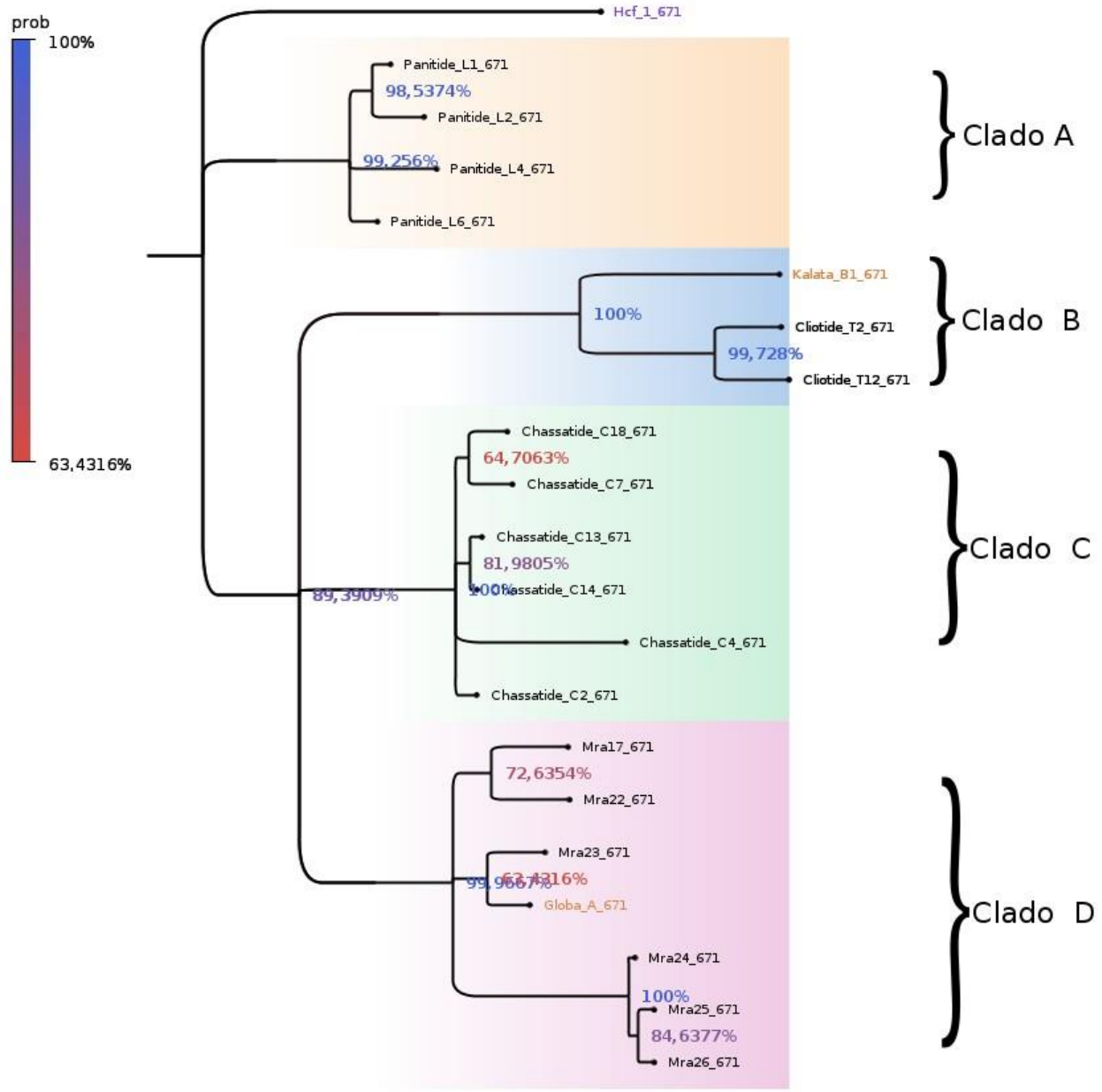


Figura 3.9: Árbol filogenetico realizado con inferencia bayesiana, analizando los datos de forma separada solo co. N de generaciones =25000000, burnin = 25 %

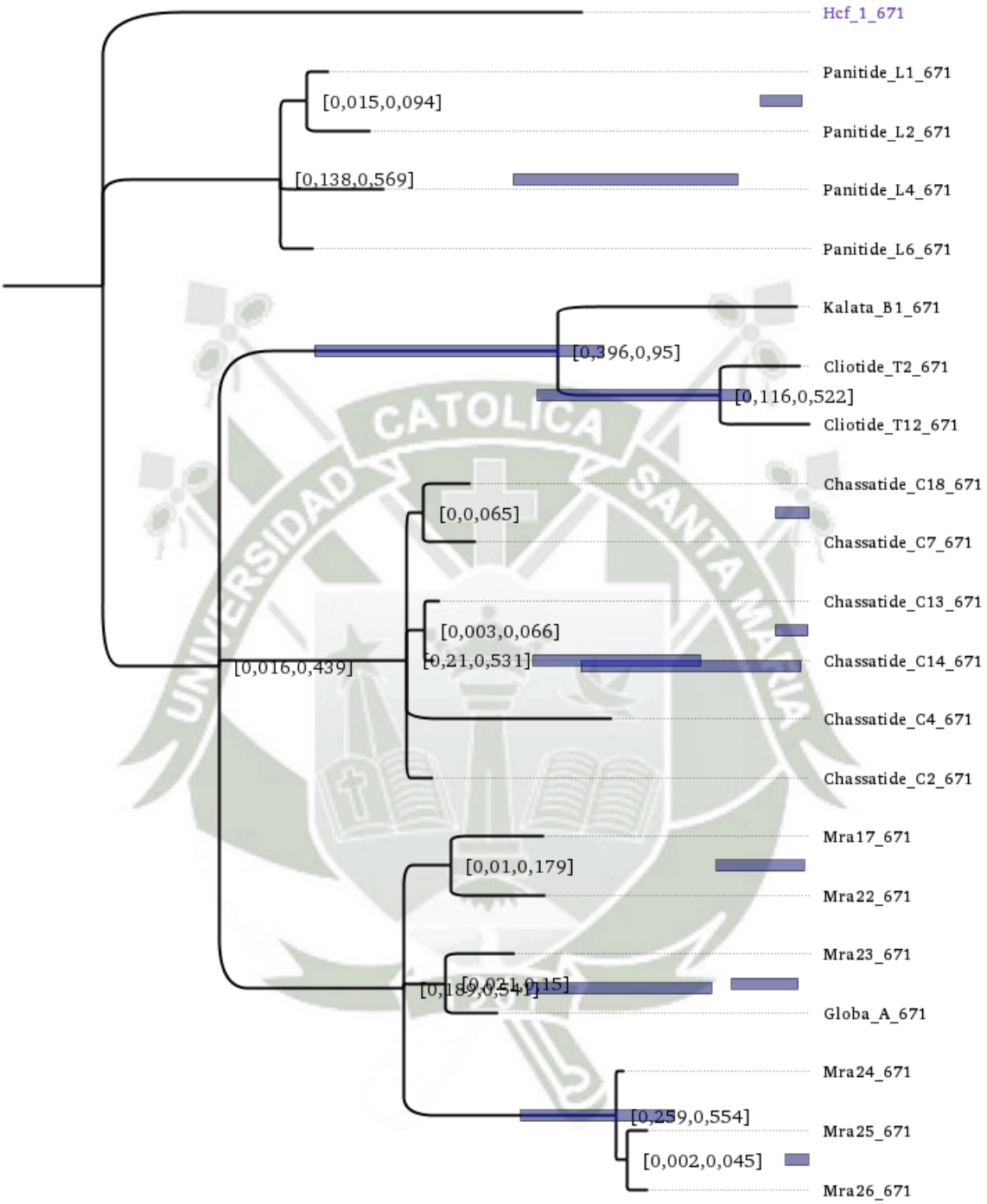
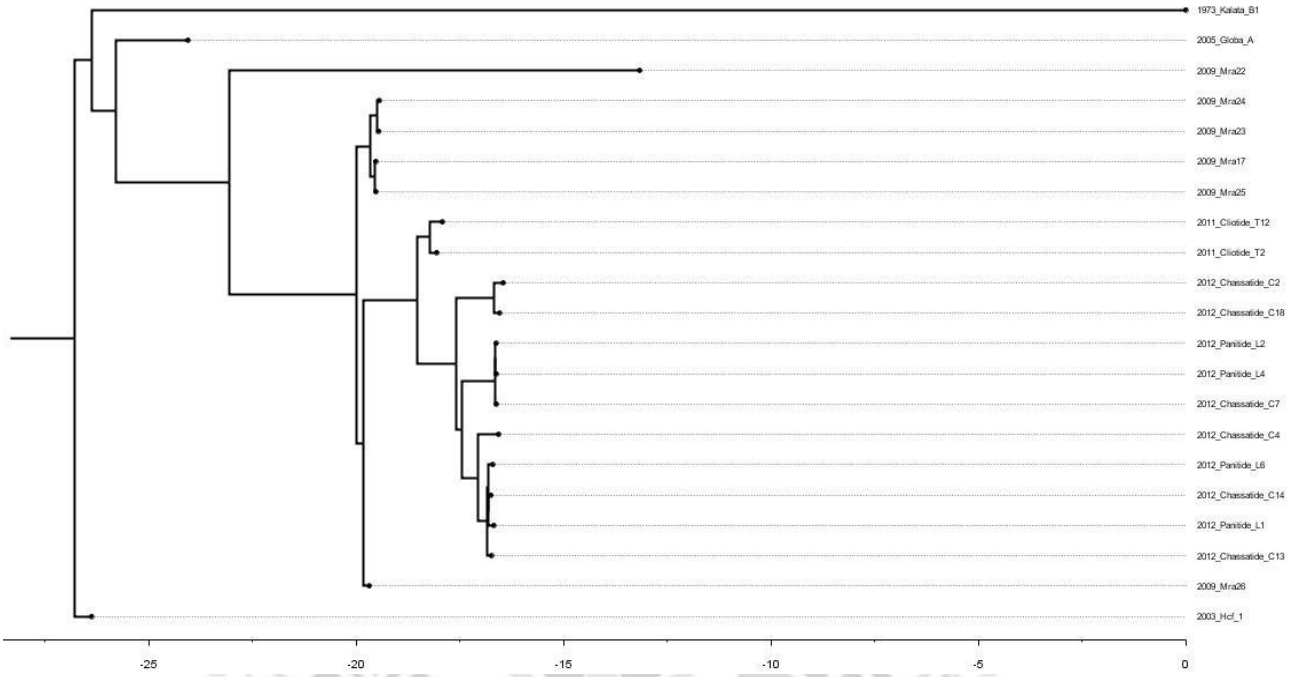


Figura 3.10: Árbol filogenetico realizado con inferencia bayesiana, analizando los valores superiores e inferiores dentro de la densidad de probabilidad más alta al 95 %

Figura 3.11: Estimación de tiempo de divergencia en millones de años sin calibración de fósiles usando un reloj molecular relajado



Los resultados para el reloj molecular relajado, mostraron una gran diferencia usando los diferentes parámetros para los árboles, Se analizó el tamaño de muestra efectiva (ESS) dándonos resultados muy diferentes para cada parámetro de árbol, mientras que para el tamaño de población constante(CS) el ESS en posterior salió un valor de 4, lo que indicaría que muchas de las muestras están correlacionadas y no siguen este parámetro, para el crecimiento exponencial (EG) salieron valore de ESS mayores a 100, lo que quiere decir que se ha alcanzado un adecuado número de muestras independientes. De la misma forma para la topología de los árboles (figura 3.11), con una distribución de clados muy diferentes en los dos parámetros, confirmando la teoría de que los ciclótidos pueden cumplir en una forma muy ligera un reloj molecular relajado con crecimiento exponencial, esto se puede deber a que cada cambio mutacional tiene que ver con los factores externos de la planta, los ciclótidos se van expresando según la planta genera algún tipo de estrés y como ya se dijo anteriormente los principales cambios dentro de la secuencia están vinculados a la funcionalidad.

Capítulo 4

Conclusiones

1. Se seleccionó las secuencias de precursores ciclotídicos a partir de bases de datos curadas, obteniendo así 23 secuencias nucleotídicas de ARNm, seleccionando solo aquellas que cumplían con los parámetros de tamaño. Después se llevó a cabo el alineamiento múltiple de las secuencias en formato FASTA mediante el método ClustalW, con un resultado que consta de un menor número de gaps y más sitios conservados.

2. Se confirmó la continuidad del análisis filogenético a través del índice de saturación, que indicó que existe cierta saturación propia de la naturaleza de los ciclótidos. El modelo de evolución apropiado es el GTR+G.

3. El análisis de las cadenas MCMC a través de la inferencia bayesiana, mostró estacionariedad y convergencia, esto por los parámetros estadísticos óptimos.

4. Se reconstruyó el árbol filogenético usando el método de inferencia bayesiana, estableciendo que los ciclótidos que están relacionadas filogenéticamente comparten una función común aún no probada. En el caso del Clado B se ve a la familia *fabaceae* y *rubiaceae* unidas con un porcentaje de probabilidad de un 100 % y en el caso del Clado D al ser una familia homóloga de *violaceae* se nota un porcentaje de probabilidad muy bajo esto por qué comparten funciones diferentes entre sí. Los ciclótidos no cumplen con un reloj molecular relajado, debido a que su síntesis está estrechamente relacionado con el rol de defensa en la planta frente a factores externos

Capítulo 5

Recomendaciones

1. Tenerse en cuenta el tamaño de las secuencias a analizar y verificar que realmente necesita una estudio como el de inferencia bayesiana.
2. Tener un conocimiento amplio sobre los programas a usar en análisis sobre filogenética.
3. Realizar futuras investigaciones sobre los ciclótidos presentes en especies endémicas peruanas.

Referencias Bibliográficas

[1] Craik DJ, editor. Advances in Botanical Research: Plant Cyclotides. vol. 76; 2015.

[2] Gunasekera S, Daly NL, Anderson MA, Craik DJ. Chemical synthesis and biosynthesis of the cyclotide family of circular proteins. IUBMB life. 2006;58(9):515–524.

[3] Craik DJ, Mylne JS, Daly NL. Cyclotides: macrocyclic peptides with applications in drug design and agriculture. Cellular and molecular life sciences. 2010;67(1):9–16.

[4] Wilkinson M, McInerney JO, Hirt RP, Foster PG, Embley TM. Of clades and clans: terms for phylogenetic relationships in unrooted trees. Trends in ecology & evolution. 2007;22(3):114–115.

[5] Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y. Predicting function: from genes to genomes and back. Journal of molecular biology. 1998;283(4):707–725.

[6] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004;32(5):1792–1797.

[7] Giribet G, Wheeler WC. On gaps. Molecular phylogenetics and evolution. 1999;13(1):132–143.

[8] Posada D. jModelTest: phylogenetic model averaging. Molecular biology and evolution. 2008;25(7):1253–1256.

[9] Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 2005;21(9):2104–2105.

[10] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution. 1987;4(4):406–425.

[11] Pelegrini PB, Quirino BF, Franco OL. Plant cyclotides: an unusual class of defense compounds. peptides. 2007;28(7):1475–1481.

[12] Gran L, Sandberg F, Sletten K. Oldenlandia affinis (R&S) DC: a plant containing uteroactive peptides used in African traditional medicine. Journal of ethnopharmacology. 2000;70(3):197–203.

[13] Saether O, Craik DJ, Campbell ID, Sletten K, Juul J, Norman DG. Elucidation of the primary and three-dimensional structure of the uterotonic polypeptide kalata B1. Biochemistry. 1995;34(13):4147–4158.

[14] Gruber CW. Global cyclotide adventure: a journey dedicated to the discovery of circular peptides from flowering plants. Peptide Science. 2010;94(5):565–572.

[15] Colgrave ML, Poth AG, Kaas Q, Craik DJ. A new “era” for cyclotide sequencing. Peptide Science. 2010;94(5):592–601.

[16] Craik DJ, Daly NL, Bond T, Waine C. Plant cyclotides: a unique family of cyclic and knotted proteins that defines the cyclic cystine knot structural motif. Journal of molecular biology. 1999;294(5):1327–1336.

[17] Felizmenio-Quimio ME, Daly NL, Craik DJ. Circular Proteins in Plants SOLUTION STRUCTURE OF A NOVEL MACROCYCLIC TRYPSIN INHIBITOR FROMMOMORDICA COCHINCHINENSIS. Journal of Biological Chemistry. 2001;276(25):22875–22882.

[18] Kaas Q, Craik DJ. Analysis and classification of circular proteins in CyBase. Peptide Science. 2010;94(5):584–591.

[19] Zhang J, Liao B, Craik DJ, Li JT, Hu M, Shu WS. Identification of two suites of cyclotide precursor genes from metallophyte Viola baoshanensis: cDNA sequence variation, alternative RNA splicing and potential cyclotide diversity. Gene. 2009;431(1):23–32.

[20] Taiz L. The plant vacuole. Journal of Experimental Biology. 1992;172(1):113–122.

[21] Conlan BF, Gillon AD, Barbeta BL, Anderson MA. Subcellular targeting and biosynthesis of cyclotides in plant cells. American journal of botany. 2011;98(12):2018–2026.

[22] Gillon AD, Saska I, Jennings CV, Guarino RF, Craik DJ, Anderson MA. Biosynthesis of circular proteins in plants. The Plant Journal. 2008;53(3):505–515.

[23] Tam JP, Wong CT. Chemical synthesis of circular proteins. Journal of Biological Chemistry. 2012;287(32):27020–27025.

[24] Zheng JS, Chang HN, Shi J, Liu L. Chemical synthesis of a cyclotide via intramolecular cyclization of peptide O-esters. Science China Chemistry. 2012;55(1):64–69.

[25] Sancheti H, Camarero JA. "Splicing up" drug discovery.: Cell-based expression and screening of genetically-encoded libraries of backbone-cyclized polypeptides. Advanced drug delivery reviews. 2009;61(11):908–917.

[26] Zheng JS, Tang S, Guo Y, Chang HN, Liu L. Synthesis of cyclic peptides and cyclic proteins via ligation of peptide hydrazides. ChemBioChem. 2012;13(4):542–546.

[27] Craik DJ, Fairlie DP, Liras S, Price D. The future of peptide-based drugs. Chemical biology & drug design. 2013;81(1):136–147.

[28] Craik D, Simonsen S, Daly N. The cyclotides: novel macrocyclic peptides as scaffolds in drug design. Current opinion in drug discovery & development. 2002;5(2):251–260.

[29] Craik DJ, Clark RJ, Daly NL. Potential therapeutic applications of the cyclotides and related cystine knot mini-proteins. Expert opinion on investigational drugs. 2007;16(5):595–604.

[30] Ireland DC, Wang CK, Wilson JA, Gustafson KR, Craik DJ. Cyclotides as natural anti-HIV agents. Peptide Science. 2008;90(1):51–60.

[31] Gunasekera S, Foley FM, Clark RJ, Sando L, Fabri LJ, Craik DJ, et al. Engineering stabilized vascular endothelial growth factor-A antagonists: synthesis, structural characterization, and bioactivity of grafted analogues of cyclotides. Journal of medicinal chemistry. 2008;51(24):7697–7704.

[32] Thongyoo P, Bonomelli C, Leatherbarrow RJ, Tate EW. Potent inhibitors of β-tryptase and human leukocyte elastase based on the MCoTI-II scaffold. Journal of medicinal chemistry. 2009;52(20):6197–6200.

[33] Jennings C, West J, Waine C, Craik D, Anderson M. Biosynthesis and insecticidal properties of plant cyclotides: the cyclic knotted proteins from Oldenlandia affinis. Proceedings of the National Academy of Sciences. 2001;98(19):10614–10619.

[34] León B. Violaceae endémicas del Perú. Revista peruana de biología. 2006;13(2):677–678.

[35] Simonsen SM, Sando L, Ireland DC, Colgrave ML, Bharathi R, Göransson U, et al. A continent of plant defense peptide diversity: cyclotides in Australian Hybanthus (Violaceae). The Plant Cell. 2005;17(11):3176–3189.

[36] Burman R, Gruber CW, Rizzardi K, Herrmann A, Craik DJ, Gupta MP, et al. Cyclotide proteins and precursors from the genus Gloeospermum: filling a blank spot in the cyclotide map of Violaceae. Phytochemistry. 2010;71(1):13–20.

[37] Craik DJ, Daly NL, Mulvenna J, Plan MR, Trabi M. Discovery, structure and biological activities of the cyclotides. Current Protein and Peptide Science. 2004;5(5):297–315.

[38] Gustafson KR, Sowder RC, Henderson LE, Parsons IC, Kashman Y, Cardellina JH, et al. Circulins A and B. Novel human immunodeficiency virus (HIV)-inhibitory macrocyclic peptides from the tropical tree Chassalia parvifolia. Journal of the American Chemical Society. 1994;116(20):9337–9338.

[39] Poth AG, Colgrave ML, Philip R, Kerenga B, Daly NL, Anderson MA, et al. Discovery of cyclotides in the Fabaceae plant family provides new insights into the cyclization, evolution, and distribution of circular proteins. ACS chemical biology. 2011;6(4):345–355.

[40] Poth AG, Mylne JS, Grassl J, Lyons RE, Millar AH, Colgrave ML, et al. Cyclotides associate with leaf vasculature and are the products of a novel precursor in petunia (Solanaceae). Journal of Biological Chemistry. 2012;287(32):27033–27046.

[41] Nguyen GKT, Lian Y, Pang EWH, Nguyen PQT, Tran TD, Tam JP. Discovery of linear cyclotides in monocot plant Panicum laxum of Poaceae family provides new insights into evolution and distribution of cyclotides in plants. Journal of Biological Chemistry. 2013;288(5):3370–3380.

[42] Suryawanshi R, Patil C, Borase H, Narkhede C, Patil S. Screening of Rubiaceae and Apocynaceae extracts for mosquito larvicidal potential. Natural product research. 2015;29(4):353–358.

[43] Camarero JA. Optimizing the future for biotechnology therapies, the key role of protein engineering. Advanced drug delivery reviews. 2009;61(11):897–898.

[44] Dörnenburg H. Cyclotide synthesis and supply: from plant to bioprocess. Peptide Science. 2010;94(5):602–610.

[45] Camarero JA, Kimura RH, Woo YH, Shekhtman A, Cantor J. Biosynthesis of a fully functional cyclotide inside living bacterial cells. Chembiochem. 2007;8(12):1363–1366.

[46] Pinto MF, Fensterseifer IC, Franco OL. Plant cyclotides: an unusual protein family with multiple functions. Springer; 2012.

[47] Felsenstein J, Felenstein J. Inferring phylogenies. vol. 2. Sinauer Associates Sunderland; 2004.

[48] Darwin C. On the origins of species by means of natural selection. vol. 247; 1859.

[49] Peña C. Métodos de inferencia filogenética. Revista Peruana de Biología. 2011;18(2):265–267.

[50] Smith AB. Rooting molecular trees: problems and strategies. Biological Journal of the Linnean Society. 1994;51(3):279–292.

[51] Huelsenbeck JP, Bollback JP, Levine AM. Inferring the root of a phylogenetic tree. Systematic biology. 2002;51(1):32–43.

[52] Landegren U, Nilsson M, Kwok PY. Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. Genome Research. 1998;8(8):769–776.

[53] Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, et al. Accessing genetic information with high-density DNA arrays. Science. 1996;p. 610–614.

[54] Crick FH. The origin of the genetic code. Journal of molecular biology. 1968;38(3):367–379.

[55] Eigen M, Gardiner W, Schuster P, Winkler-Oswatitsch R. The origin of genetic information. Scientific american. 1981;244(4):88–119.

[56] Salemi M, Lemey P, Vandamme AM. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. Cambridge University Press; 2009. Available from: https://books.google.com.pe/books?id=DeD_lQ-kBPQC.

[57] Bremer K. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution. 1988;42(4):795–803.

[58] Cavalli-Sforza LL, Edwards AW. Phylogenetic analysis: models and estimation procedures. Evolution. 1967;21(3):550–570.

[59] Lande R. Natural selection and random genetic drift in phenotypic evolution. Evolution. 1976;30(2):314–334.

[60] Nei M, Maruyama T, Chakraborty R. The bottleneck effect and genetic variability in populations. Evolution. 1975;29(1):1–10.

[61] Mayr E. Populations, species, and evolution: an abridgment of animal species and evolution. Harvard University Press; 1970.

[62] Elena SF, Lenski RE. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nature reviews Genetics. 2003;4(6):457.

[63] Hegreness M, Shoresh N, Hartl D, Kishony R. An equivalence principle for the incorporation of favorable mutations in asexual populations. Science. 2006;311(5767):1615–1617.

[64] Hartl DL, Clark AG, Clark AG. Principles of population genetics. vol. 116. Sinauer associates Sunderland; 1997.

[65] Dudek RW. High-yield Cell and Molecular Biology. No. v. 845 in High-yield Cell and Molecular Biology. Wolters Kluwer/Lippincott Williams & Wilkins; 2007. Available from: https://books.google.com.pe/books?id=g-d--DOdnQAC.

[66] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols. 2009;4(7):1073–1081.

[67] Strachan T, Read AP. Human Molecular Genetics 3. Garland Science; 2004. Available from: https://books.google.com.pe/books?id=g4hC63UrPbUC.

[68] Freese EB. Transitions and transversions induced by depurinating agents. Proceedings of the National Academy of Sciences. 1961;47(4):540–545.

[69] Broughton RE, Stanley SE, Durrett RT. Quantification of homoplasy for nucleotide transitions and transversions and a reexamination of assumptions in weighted phylogenetic analysis. Systematic Biology. 2000;49(4):617–627.

[70] Wakeley J. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. Trends in ecology & evolution. 1996;11(4):158–162.

[71] Solari AJ. Genética humana: fundamentos y aplicaciones en medicina. Médica Panamericana; 2004. Available from: https://books.google.com.pe/books?id= e-slX7S1KdsC.

[72] Wheeler WC, Aagesen L, Arango CP, Faivovich J, Grant T, D'Haese C, et al. Dynamic homology and phylogenetic systematics: a unified approach using POY. American Museum of Natural History; 2006.

[73] Remane J. The concept of homology in phylogenetic research—its meaning and possible applications. Paläontologische Zeitschrift. 1983;57(3):267–269.

[74] Campbell NA, Reece JB. Biología. Editorial Medica Panamericana Sa de; 2007. Available from: https://books.google.com.pe/books?id=QcU0yde9PtkC.

[75] Zhou J. Microbial functional genomics. John Wiley & Sons; 2004.

[76] Kimura M. The neutral theory of molecular evolution: a review of recent evidence. Revista de la genética. 1991;66(4):367–386.

[77] Kimura M. The Neutral Theory of Molecular Evolution. Cambridge University Press; 1983. Available from: https://books.google.es/books?id=olIoSumPevYC.

[78] Takahata N. Neutral theory of molecular evolution. Current opinion in genetics & development. 1996;6(6):767–772.

[79] Thorpe JP. The molecular clock hypothesis: biochemical evolution, genetic differentiation and systematics. Annual Review of Ecology and Systematics. 1982;13:139–168.

[80] Edgar RC, Batzoglou S. Multiple sequence alignment. Current opinion in structural biology. 2006;16(3):368–373.

[81] Higgins DG, Sharp PM. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene. 1988;73(1):237–244.

[82] Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology. 2000;302(1):205–217.

[83] Simmons MP, Ochoterena H. Gaps as characters in sequence-based phylogenetic analyses. Systematic biology. 2000;49(2):369–381.

[84] Thorne JL. Models of protein sequence evolution and their applications. Current opinion in genetics & development. 2000;10(6):602–605.

[85] Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nature methods. 2012;9(8):772–772.

[86] Sugiura N. Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. Communications in Statistics-Theory and Methods. 1978;7(1):13–26.

[87] Schwarz G. Estimating the dimension of a model Ann Stat 6: 461–464. Find this article online. 1978;.

[88] Sokal RR. A statistical method for evaluating systematic relationships. Univ Kans Sci Bull. 1958;38:1409–1438.

[89] Stefan Van Dongen T, Winnepenninckx B. Multiple UPGMA and neighbor-joining trees and the performance of some computer packages. Mol Biol Evol. 1996;13(2):309–313.

[90] Bayer U. Pattern Recognition Problems in Geology and Paleontology. 1985;.

[91] Kovach W. MVSP Plus: multivariate statistical package, version 2.1. Kovach Computing Services, Pentraeth, Anglesey. 1993;.

[92] Sneath PH, Sokal RR, et al. Numerical taxonomy. The principles and practice of numerical classification.; 1973.

[93] Drummond A, Rodrigo AG. Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. Molecular Biology and Evolution. 2000;17(12):1807–1815.

[94] Wu X, Wan XF, Wu G, Xu D, Lin G. Phylogenetic analysis using complete signature information of whole genomes and clustered Neighbour-Joining method. International journal of bioinformatics research and applications. 2006;2(3):219–248.

[95] Källersjö M, Farris JS, Kluge AG, Bult C. Skewness and permutation. Cladistics. 1992;8(3):275–287.

[96] Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(30):11030–11035.

[97] Kidd KK, Sgaramella-Zonta LA. Phylogenetic analysis: concepts and methods. American journal of human genetics. 1971;23(3):235.

[98] Rzhetsky A, Nei M. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Molecular biology and evolution. 1993;10(5):1073–1095.

[99] Sourdis J, Nei M. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. Molecular biology and evolution. 1988;5(3):298–311.

[100] Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. Heterotachy and long-branch attraction in phylogenetics. BMC evolutionary biology. 2005;5(1):50.

[101] Kim J. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. Systematic Biology. 1996;45(3):363–374.

[102] Takezaki N, Nei M. Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant. Journal of molecular evolution. 1994;39(2):210–218.

[103] Zharkikh A, Li WH. Inconsistency of the maximum-parsimony method: the case of five taxa with a molecular clock. Systematic Biology. 1993;42(2):113–125.

[104] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution. 1981;17(6):368–376.

[105] Hasegawa M, Kishino H, Saitou N. On the maximum likelihood method in molecular phylogenetics. Journal of molecular evolution. 1991;32(5):443–445.

[106] Yang Z. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular biology and evolution. 1993;10(6):1396–1401.

[107] Ronquist F. Bayesian inference of character evolution. Trends in ecology & evolution. 2004;19(9):475–481.

[108] Huelsenbeck JP, Larget B, Miller RE, Ronquist F. Potential applications and pitfalls of Bayesian inference of phylogeny. Systematic biology. 2002;51(5):673–688.

[109] Felsenstein J. Statistical inference and the estimation of phylogenies. University of Chicago, Department of Zoology; 1968.

[110] Archibald JK, Mort ME, Crawford DJ. Bayesian inference of phylogeny: a non-technical primer. Taxon. 2003;52(2):187–191.

[111] Box GE, Tiao GC. Bayesian inference in statistical analysis. vol. 40. John Wiley & Sons; 2011.

[112] Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. Bayesian inference of phylogeny and its impact on evolutionary biology. science. 2001;294(5550):2310–2314.

[113] Mau B, Newton MA, Larget B. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics. 1999;55(1):1–12.

[114] Geyer CJ. Practical markov chain monte carlo. Statistical Science. 1992;p. 473–483.

[115] Pagel M, Meade A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. The American Naturalist. 2006;167(6):808–825.

[116] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. The journal of chemical physics. 1953;21(6):1087–1092.

[117] Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970;57(1):97–109.

[118] Chib S, Greenberg E. Understanding the metropolis-hastings algorithm. The american statistician. 1995;49(4):327–335.

[119] Kuhner MK, Yamato J, Felsenstein J. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics. 1995;140(4):1421–1430.

[120] Xia X, Xie Z. DAMBE: software package for data analysis in molecular biology and evolution. Journal of heredity. 2001;92(4):371–373.

[121] Okonechnikov K, Golosova O, Fursov M, team U. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012;28(8):1166–1167.

[122] Maddison WP, Maddison DR. Mesquite: a modular system for evolutionary analysis. 2001;.

[123] Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17(8):754–755.

[124] Rambaut A, Drummond A. FigTree. Program distributed by the authors; 2013.

[125] Rambaut A, Suchard MA, Xie D, Drummond AJ. Tracer v1. 6. 2014. 2015;.

[126] Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology. 2007;7(1):214.

[127] Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular biology and evolution. 2012;29(8):1969–1973.

[128] Nguyen GKT, Zhang S, Nguyen NTK, Nguyen PQT, Chiu MS, Hardjojo A, et al. Discovery and characterization of novel cyclotides originated from chimeric precursors consisting of albumin-1 chain a and cyclotide domains in the Fabaceae family. Journal of Biological Chemistry. 2011;286(27):24275–24287.

[129] Pelegrini PB, Franco OL. Plant γ-thionins: novel insights on the mechanism of action of a multi-functional class of defense proteins. The international journal of biochemistry & cell biology. 2005;37(11):2239–2253.

Parte I

Apéndices

.1. Anexo 1

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
Cycloviolacin 01	*Violaceae*	Bracalet	17433722	30
Vhr 1	*Violaceae*	Bracalet	152031727	30
Tricyclon A	*Violaceae*	Bracalet	67463910	33
Cycloviolacin O2	*Violaceae*	Bracalet	190358835	30
Cycloviolacin O8	*Violaceae*	Bracalet	CyBase	31
cycloviolacin O11	*Violaceae*	Bracalet	Cybase	31
vodo N	*Violaceae*	Möbius	47117393	29
cycloviolacin H1	*Violaceae*	Bracalet	17865458	30
cycloviolacin O9	*Violaceae*	Bracalet	17433005	30
vico A	*Violaceae*	Bracalet	Cybase	31
vitri A	*Violaceae*	Bracalet	47117394	30
cycloviolacin O12	*Violaceae*	Möbius	46577590	29
vodo N	*Violaceae*	Möbius	47117392	29
vico B	*Violaceae*	Bracalet	76365060	31
Hypa A	*Violaceae*	Bracalet	17433009	30
cycloviolacin O4	*Violaceae*	Bracalet	152031585	30
cycloviolacin O3	*Violaceae*	Bracalet	17432999	30

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
cycloviolacin O5	*Violaceae*	Bracalet	17433001	30
cycloviolacin O6	*Violaceae*	Bracalet	17433002	31
cycloviolacin O7	*Violaceae*	Bracalet	17433003	30
cycloviolacin O10	*Violaceae*	Bracalet	17433006	30
varv peptide B	*Violaceae*	Möbius	17433211	30
varv peptide C	*Violaceae*	Möbius	17433212	31
varv peptide D	*Violaceae*	Möbius	17433213	30
varv peptide F	*Violaceae*	Möbius	17433215	30
varv peptide G	*Violaceae*	Möbius	17433216	30
varv peptide H	*Violaceae*	Möbius	17433217	30
cycloviolin A	*Violaceae*	Bracalet	76364163	31
cycloviolin B	*Violaceae*	Bracalet	76365059	28
cycloviolin C	*Violaceae*	Bracalet	76365061	30
cycloviolin D	*Violaceae*	Bracalet	76365062	30
Violapeptide 1	*Violaceae*	Möbius	Cybase	29
Vhl-1	*Violaceae*	Bracalet	73620924	31
Vhl-2	*Violaceae*	Möbius	158562877	30

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
cycloviolacin H3	*Violaceae*	Möbius	158562878	30
cycloviolacin H2	*Violaceae*	Bracalet	158562879	29
Hyfl A	*Violaceae*	Bracalet	82592912	31
Hyfl B	*Violaceae*	Möbius	82592913	32
Hyfl C	*Violaceae*	Möbius	82592914	32
Hyfl D	*Violaceae*	——	Cybase	31
Hyfl E	*Violaceae*	——	Cybase	29
Hyfl F	*Violaceae*	——	Cybase	30
Hyfl I	*Violaceae*	——	Cybase	30
Hyfl J	*Violaceae*	——	Cybase	28
Hyfl K	*Violaceae*	——	Cybase	30
Hyfl L	*Violaceae*	——	Cybase	30
Hyfl M	*Violaceae*	——	Cybase	29
tricyclon B	*Violaceae*	——	Cybase	33
cycloviolacin H4	*Violaceae*	Bracalet	158562880	30
cycloviolacin O13	*Violaceae*	Bracalet	Cybase	30
violacin A	*Violaceae*	——	88193106	27

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
cycloviolacin O14	*Violaceae*	Möbius	152013458	31
cycloviolacin O15	*Violaceae*	Möbius	152013459	29
cycloviolacin O16	*Violaceae*	Möbius	152013460	29
cycloviolacin O17	*Violaceae*	Bracalet	152013461	30
cycloviolacin O18	*Violaceae*	Bracalet	152013462	30
cycloviolacin O19	*Violaceae*	Bracalet	152013463	31
cycloviolacin O20	*Violaceae*	Bracalet	152013464	30
cycloviolacin O21	*Violaceae*	Möbius	152013465	29
cycloviolacin O22	*Violaceae*	Möbius	152013466	29
cycloviolacin O23	*Violaceae*	Möbius	152013467	31
cycloviolacin O24	*Violaceae*	Möbius	152013468	30
cycloviolacin O25	*Violaceae*	Bracalet	152013469	31
cycloviolacin Y1	*Violaceae*	——	Cybase	33
cycloviolacin Y2	*Violaceae*	——	Cybase	33
cycloviolacin Y3	*Violaceae*	——	Cybase	33
cycloviolacin Y4	*Violaceae*	——	Cybase	30
cycloviolacin Y5	*Violaceae*	——	Cybase	30

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
vibi A	*Violaceae*	Möbius	190359078	29
vibi B	*Violaceae*	Möbius	190359079	29
vibi C	*Violaceae*	Möbius	190359080	29
vibi D	*Violaceae*	Möbius	190359081	29
vibi E	*Violaceae*	——	Cybase	30
vibi F	*Violaceae*	Bracalet	190359083	31
vibi G	*Violaceae*	Bracalet	190359084	31
vibi H	*Violaceae*	Bracalet	190359085	31
vibi I	*Violaceae*	——	Cybase	30
vibi J	*Violaceae*	——	Cybase	31
vibi K	*Violaceae*	——	Cybase	30
Viba 2	*Violaceae*	——	Cybase	30
Viba 5	*Violaceae*	——	Cybase	30
Viba 10	*Violaceae*	——	Cybase	29
Viba 12	*Violaceae*	——	Cybase	30
Viba 14	*Violaceae*	——	Cybase	34
Viba 17	*Violaceae*	——	Cybase	29

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
Viba 15	*Violaceae*	——	Cybase	29
Mram 1	*Violaceae*	——	Cybase	31
Mram 2	*Violaceae*	——	Cybase	30
Mram 3	*Violaceae*	——	Cybase	29
Mram 4	*Violaceae*	——	Cybase	31
Mram 5	*Violaceae*	——	Cybase	31
Mram 6	*Violaceae*	——	Cybase	31
Mram 7	*Violaceae*	——	Cybase	31
Mram 8	*Violaceae*	——	Cybase	30
Mram 9	*Violaceae*	——	Cybase	30
Mram 10	*Violaceae*	——	Cybase	31
Mram 11	*Violaceae*	——	Cybase	29
Mram 12	*Violaceae*	——	Cybase	30
Mram 13	*Violaceae*	——	Cybase	29
Mram 14	*Violaceae*	——	Cybase	31
Viba 1	*Violaceae*	——	Cybase	30
Viba 3	*Violaceae*	——	Cybase	30

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
Viba 4	*Violaceae*	——	Cybase	30
Viba 6	*Violaceae*	——	Cybase	30
Viba 7	*Violaceae*	——	Cybase	31
Viba 8	*Violaceae*	——	Cybase	30
Viba 9	*Violaceae*	——	Cybase	30
Viba 11	*Violaceae*	——	Cybase	30
Viba 13	*Violaceae*	——	Cybase	30
Viba 16	*Violaceae*	——	Cybase	29
Vpl-1	*Violaceae*	——	Cybase	31
Vpf-1	*Violaceae*	——	Cybase	30
cO31	*Violaceae*	——	Cybase	29
cO28	*Violaceae*	——	Cybase	29
cO32	*Violaceae*	——	Cybase	30
cO33	*Violaceae*	——	Cybase	29
cO34	*Violaceae*	——	Cybase	29
cO35	*Violaceae*	——	Cybase	29
cO29	*Violaceae*	——	Cybase	30

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
cO30	*Violaceae*	——	Cybase	30
cO26	*Violaceae*	——	Cybase	31
cO27	*Violaceae*	——	Cybase	31
Globa F	*Violaceae*	——	Cybase	31
Globa A	*Violaceae*	——	Cybase	30
Globa B	*Violaceae*	——	Cybase	31
Globa D	*Violaceae*	——	Cybase	30
Globa E	*Violaceae*	——	Cybase	30
Globa C	*Violaceae*	——	Cybase	29
Glopa D	*Violaceae*	——	Cybase	31
Glopa E	*Violaceae*	——	Cybase	30
Glopa A	*Violaceae*	——	Cybase	32
Glopa B	*Violaceae*	——	Cybase	32
Glopa C	*Violaceae*	——	Cybase	31
Co36	*Violaceae*	——	Cybase	30
cycloviolacin T1	*Violaceae*	——	Cybase	29
vaby A	*Violaceae*	——	Cybase	29

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
vaby B	*Violaceae*	——	Cybase	29
vaby C	*Violaceae*	——	Cybase	29
vaby D	*Violaceae*	——	Cybase	30
vaby E	*Violaceae*	——	Cybase	30
vitri B	*Violaceae*	——	Cybase	29
vitri C	*Violaceae*	——	Cybase	29
vitri D	*Violaceae*	——	Cybase	29
vitri E	*Violaceae*	——	Cybase	29
vitri F	*Violaceae*	——	Cybase	31
viphi A	*Violaceae*	——	Cybase	31
viphi B	*Violaceae*	——	Cybase	29
viphi C	*Violaceae*	——	Cybase	30
viphi D	*Violaceae*	——	Cybase	30
viphi E	*Violaceae*	——	Cybase	31
viphi F	*Violaceae*	——	Cybase	31
viphi G	*Violaceae*	——	Cybase	31
viphi H	*Violaceae*	——	Cybase	30

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
Vigno 1	*Violaceae*	——	Cybase	29
Vigno 2	*Violaceae*	——	Cybase	30
Vigno 3	*Violaceae*	——	Cybase	29
Vigno 4	*Violaceae*	——	Cybase	29
Vigno 5	*Violaceae*	——	Cybase	29
Vigno 6	*Violaceae*	——	Cybase	31
Vigno 7	*Violaceae*	——	Cybase	31
Vigno 8	*Violaceae*	——	Cybase	30
Vigno 9	*Violaceae*	——	Cybase	30
Vigno 10	*Violaceae*	——	Cybase	31
vocC	*Violaceae*	——	Cybase	29
kalata B1	*Rubiaceae*	Möbius	253722944	29
kalata B2	*Rubiaceae*	Möbius	17433722	29
palicourein	*Rubiaceae*	Bracelet	84027829	37
circulin A	*Rubiaceae*	Bracelet	17433719	30
kalata B6	*Rubiaceae*	Möbius	Cybase	30
kalata B3	*Rubiaceae*	Möbius	Cybase	30

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
kalata B7	*Rubiaceae*	Möbius	558704667	29
kalata B4	*Rubiaceae*	Möbius	152031627	29
cyclopsychotride A	*Rubiaceae*	Bracelet	17433720	31
kalata S	*Rubiaceae*	Möbius	17433131	29
circulin B	*Rubiaceae*	Bracelet	17433721	31
circulin C	*Rubiaceae*	Bracelet	76364211	30
circulin D	*Rubiaceae*	Bracelet	76364212	30
circulin E	*Rubiaceae*	Bracelet	76364213	30
circulin F	*Rubiaceae*	Bracelet	76364214	29
Kalata-B5	*Rubiaceae*	Bracelet	254763307	30
Hcf-1	*Rubiaceae*	——	Cybase	30
Htf-1	*Rubiaceae*	——	Cybase	30
kalata B8	*Rubiaceae*	Bracelet	152032551	31
kalata B9	*Rubiaceae*	Bracelet	152032552	31
kalata B9 linear	*Rubiaceae*	Bracelet	Cybase	31
kalata B10	*Rubiaceae*	Möbius	152032543	30
kalata B10 linear	*Rubiaceae*	Möbius	Cybase	30

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
kalata B11	*Rubiaceae*	Möbius	152032544	29
kalata B12	*Rubiaceae*	Möbius	152032545	28
kalata B13	*Rubiaceae*	Möbius	152032546	30
kalata B14	*Rubiaceae*	Möbius	152032547	30
kalata B15	*Rubiaceae*	Möbius	152032548	29
kalata B16	*Rubiaceae*	Bracelet	152032549	30
kalata B17	*Rubiaceae*	Bracelet	152032549	30
kalata B18	*Rubiaceae*	——	Cybase	30
PS-1	*Rubiaceae*	——	Cybase	31
CD-1	*Rubiaceae*	——	Cybase	34
hcf-1 variant	*Rubiaceae*	——	Cybase	29
psyle A	*Rubiaceae*	——	Cybase	28
psyle B	*Rubiaceae*	——	Cybase	28
psyle C	*Rubiaceae*	——	Cybase	25
psyle D	*Rubiaceae*	——	Cybase	31
psyle E	*Rubiaceae*	——	Cybase	31
psyle F	*Rubiaceae*	——	Cybase	31

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
kalata B19	*Rubiaceae*	——	Cybase	30
Oak6 cyclotide 2	*Rubiaceae*	——	Cybase	30
Oak7 cyclotide	*Rubiaceae*	——	Cybase	29
Oak8 cyclotide	*Rubiaceae*	——	Cybase	30
Oak6 cyclotide 1	*Rubiaceae*	——	Cybase	30
hedyotide B1	*Rubiaceae*	——	Cybase	30
Parigidin-br1	*Rubiaceae*	Bracelet	380877061	32
hedyotide B2	*Rubiaceae*	Bracelet	Cybase	29
Caripe 1	*Rubiaceae*	——	Cybase	31
Caripe 2	*Rubiaceae*	——	Cybase	31
Caripe 4	*Rubiaceae*	——	Cybase	27
Caripe 6	*Rubiaceae*	——	Cybase	28
Chacur 1	*Rubiaceae*	——	Cybase	29
Psybra 1	*Rubiaceae*	——	Cybase	29
Paltet 1	*Rubiaceae*	——	Cybase	29
Psypoe 1	*Rubiaceae*	——	Cybase	29
Caripe 7	*Rubiaceae*	——	Cybase	34

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
Caripe 8	*Rubiaceae*	——	Cybase	31
chassatide C18	*Rubiaceae*	——	Cybase	30
chassatide C16	*Rubiaceae*	——	Cybase	31
chassatide C15	*Rubiaceae*	——	Cybase	31
chassatide C13	*Rubiaceae*	——	Cybase	31
chassatide C17	*Rubiaceae*	——	Cybase	29
chassatide C14	*Rubiaceae*	——	Cybase	31
chassatide C8	*Rubiaceae*	——	Cybase	30
chassatide C7	*Rubiaceae*	——	Cybase	29
chassatide C4	*Rubiaceae*	——	Cybase	29
chassatide C2	*Rubiaceae*	——	Cybase	31
chassatide C1	*Rubiaceae*	——	Cybase	29
chassatide C3	*Rubiaceae*	——	Cybase	29
chassatide C5	*Rubiaceae*	——	Cybase	31
chassatide C6	*Rubiaceae*	——	Cybase	31
chassatide C9	*Rubiaceae*	——	Cybase	30
chassatide C10	*Rubiaceae*	——	Cybase	29

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
chassatide C11	*Rubiaceae*	——	Cybase	28
chassatide C12	*Rubiaceae*	——	Cybase	28
Cter A	*Fabaceae*	Bracelet	325530015	31
Cter B	*Fabaceae*	Bracelet	353526225	31
Cter C	*Fabaceae*	Bracelet	325530017	31
Cter D	*Fabaceae*	Bracelet	325530018	31
Cter E	*Fabaceae*	Bracelet	325530019	31
Cter F	*Fabaceae*	Bracelet	325530020	30
Cter G	*Fabaceae*	Bracelet	325530021	30
Cter H	*Fabaceae*	Bracelet	325530022	30
Cter I	*Fabaceae*	——	Cybase	31
Cter J	*Fabaceae*	Bracelet	325530024	31
Cter K	*Fabaceae*	Bracelet	325530025	29
Cter L	*Fabaceae*	Bracelet	325530026	29
Cter M	*Fabaceae*		333360993	29
Cter N	*Fabaceae*	Möbius	347602402	29
Cter O	*Fabaceae*	Bracelet	347602403	30

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
Cter P	*Fabaceae*	Bracelet	347602404	30
Cter Q	*Fabaceae*	Bracelet	347602405	30
Cter R	*Fabaceae*	Bracelet	347602406	31
cliotide T8	*Fabaceae*	——	Cybase	30
cliotide T9	*Fabaceae*	——	Cybase	30
cliotide T2	*Fabaceae*	——	Cybase	30
cliotide T12	*Fabaceae*	——	Cybase	30
CPTI II	*Curcubitaceae*	Trypsin Inhibitor	6980536	29
CPTI III	*Curcubitaceae*	Trypsin Inhibitor	125014	32
CMTI IV	*Curcubitaceae*	Trypsin Inhibitor	125015	32
CMTI I	*Curcubitaceae*	Trypsin Inhibitor	125005	29
LLDTI II	*Curcubitaceae*	Trypsin Inhibitor	248752	29
CVTI I	*Curcubitaceae*	Trypsin Inhibitor	125004	30
CMCTI I	*Curcubitaceae*	Trypsin Inhibitor	257181	28
CMCTI III	*Curcubitaceae*	Trypsin Inhibitor	400076	30
CSTI IV	*Curcubitaceae*	Trypsin Inhibitor	125016	30
CSTI IIB	*Curcubitaceae*	Trypsin Inhibitor	125010	32

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
CmeTI B	*Curcubitaceae*	Trypsin Inhibitor	1246050	29
LCTI II	*Curcubitaceae*	Trypsin Inhibitor	125012	30
LCTI I	*Curcubitaceae*	Trypsin Inhibitor	125006	29
LCTI III	*Curcubitaceae*	Trypsin Inhibitor	913520	29
TGTI I	*Curcubitaceae*	Trypsin Inhibitor	547745	28
McoTI II	*Curcubitaceae*	Trypsin Inhibitor	8928147	34
McoTI I	*Curcubitaceae*	Trypsin Inhibitor	8928146	34
McoTI III	*Curcubitaceae*	Trypsin Inhibitor	8928148	30
MCTI I	*Curcubitaceae*	Trypsin Inhibitor	125007	30
MCTI III	*Curcubitaceae*	Trypsin Inhibitor	61213645	30
EETI II	*Curcubitaceae*	Trypsin Inhibitor	547744	30
BDTI II	*Curcubitaceae*	Trypsin Inhibitor	125009	29
MRTI I	*Curcubitaceae*	Trypsin Inhibitor	125008	29
MCEI IV	*Curcubitaceae*	Trypsin Inhibitor	1041919	31
MCEI III	*Curcubitaceae*	Trypsin Inhibitor	1041918	30
MCEI II	*Curcubitaceae*	Trypsin Inhibitor	1041917	29
Panitide L1	*Poaceae*	——	Cybase	28

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
Panitide L2	*Poaceae*	——	Cybase	28
Panitide L4	*Poaceae*	——	Cybase	27
Panitide L6	*Poaceae*	——	Cybase	28
Panitide L3	*Poaceae*	——	Cybase	27
Panitide L5	*Poaceae*	——	Cybase	28
Panitide L7	*Poaceae*	——	Cybase	27
Panitide L8	*Poaceae*	——	Cybase	26
Z. mays G	*Poaceae*	——	GenBank: ACG26826	29
Z. mays L	*Poaceae*	——	GenBank: ACG45070	31
Z. mays M	*Poaceae*	——	GenBank: ACG42356	31
Z. mays P	*Poaceae*	——	GenBank: ACG46325	38
Phyb A	*Solanaceae*	——	Cybase	30
Phyb D	*Solanaceae*	——	Cybase	30
Phyb E	*Solanaceae*	——	Cybase	30
Phyb F	*Solanaceae*	——	Cybase	30
Phyb G	*Solanaceae*	——	Cybase	30
Phyb H	*Solanaceae*	——	Cybase	30

Tabla 1: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de accseo y número de aminoácidos

Ciclótido	Familia	Subfamilia	Acceso	N° Aminoacidos
Phyb I	*Solanaceae*	——	Cybase	31
Phyb J	*Solanaceae*	——	Cybase	31
Phyb K	*Solanaceae*	——	Cybase	31
Phyb L	*Solanaceae*	——	Cybase	32
Phyb B	*Solanaceae*	——	Cybase	30
Phyb C	*Solanaceae*	——	Cybase	30

.2. Anexo 2

Tabla 2: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de acceso y número de aminoácidos

Ciclótido	Familia	Especie	N° Acceso	N° Nucleótidos
kalata B1	*RUBIACEAE*	*Oldenlandia affinis*	FJ211184.1	456
kalata B2	*RUBIACEAE*	*Oldenlandia affinis*	AF393828.1	993
kalata B7	*RUBIACEAE*	*Oldenlandia affinis*	AF393827.1	657
Oak6 cyclotide 2	*RUBIACEAE*	*Oldenlandia affinis*	GU250888.1	596
Oak7 cyclotide	*RUBIACEAE*	*Oldenlandia affinis*	GU250889.1	561
Oak8 cyclotide	*RUBIACEAE*	*Oldenlandia affinis*	GU250890.1	585
Caripe 2	*RUBIACEAE*	*Carapichea ipecacuanha*	KC807202.1	343
chassatide C18	*RUBIACEAE*	*Chassalia chartacea*	JQ309971.1	433
chassatide C16	*RUBIACEAE*	*Chassalia chartacea*	JQ309969.1	378
chassatide C15	*RUBIACEAE*	*Chassalia chartacea*	JQ309968.1	378
chassatide C13	*RUBIACEAE*	*Chassalia chartacea*	JQ309966.1	455
chassatide C17	*RUBIACEAE*	*Chassalia chartacea*	JQ309970.1	520
chassatide C14	*RUBIACEAE*	*Chassalia chartacea*	JQ309967.1	447
chassatide C8	*RUBIACEAE*	*Chassalia chartacea*	JQ309965.1	511
chassatide C7	*RUBIACEAE*	*Chassalia chartacea*	JQ309964.1	440
chassatide C4	*RUBIACEAE*	*Chassalia chartacea*	JQ309963.1	468
chassatide C2	*RUBIACEAE*	*Chassalia chartacea*	JQ309962.1	429

Tabla 2: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de acceso y número de aminoácidos

Ciclótido	Familia	Especie	N° Acceso	N° Nucleótidos
cycloviolacin O8	*VIOLACEAE*	*Viola odorata*	FJ211181.1	624
cycloviolacin O11	*VIOLACEAE*	*Viola odorata*	AY630563.1	615
cycloviolacin O9	*VIOLACEAE*	*Viola biflora*	EU046622.1	529
vitri A	*VIOLACEAE*	*Viola biflora*	EU046623.1	542
Hyfl D	*VIOLACEAE*	*Hybanthus floribundus subsp. floribundus*	DQ192575.1	321
Hyfl E	*VIOLACEAE*	*Hybanthus floribundus subsp. floribundus*	DQ192576.1	261
Hyfl F	*VIOLACEAE*	*Hybanthus floribundus subsp. floribundus*	DQ192577.1	108
Hyfl I	*VIOLACEAE*	*Hybanthus floribundus subsp. floribundus*	DQ187928.1	318
Hyfl J	*VIOLACEAE*	*Hybanthus floribundus subsp. floribundus*	DQ187929.1	273
Hyfl K	*VIOLACEAE*	*Hybanthus floribundus subsp. floribundus*	DQ187930.1	318
Hyfl L	*VIOLACEAE*	*Hybanthus floribundus subsp. floribundus*	DQ187931.1	318
Hyfl M	*VIOLACEAE*	*Hybanthus floribundus subsp. floribundus*	DQ187932.1	309
violacin A	*VIOLACEAE*	*Viola odorata*	DQ365813.1	588
vibi E	*VIOLACEAE*	*Viola biflora*	EU046618.1	549
vibi I	*VIOLACEAE*	*Viola biflora*	EU046619.1	535
vibi J	*VIOLACEAE*	*Viola biflora*	EU046620.1	596
vibi K	*VIOLACEAE*	*Viola biflora*	EU046621.1	553

Tabla 2: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de acceso y número de aminoácidos

Ciclótido	Familia	Especie	N° Acceso	N° Nucleótidos
Viba 2	*VIOLACEAE*	*Viola baoshanensis*	DQ851860.1	643
Viba 5	*VIOLACEAE*	*Viola baoshanensis*	DQ851861.1	676
Viba 10	*VIOLACEAE*	*Viola baoshanensis*	DQ851862.1	686
Viba 12	*VIOLACEAE*	*Viola baoshanensis*	DQ851863.1	726
Viba 14	*VIOLACEAE*	*Viola baoshanensis*	DQ851864.1	587
Mra4	*VIOLACEAE*	*Melicytus ramiflorus*	EF103478.1	788
Mra13	*VIOLACEAE*	*Melicytus ramiflorus*	EF103479.1	303
Mra14	*VIOLACEAE*	*Melicytus ramiflorus*	EF103465.1	672
Mra17	*VIOLACEAE*	*Melicytus ramiflorus*	EF103467.1	433
Mra29	*VIOLACEAE*	*Melicytus ramiflorus*	EF103476.1	660
Mra30	*VIOLACEAE*	*Melicytus ramiflorus*	EF103477.1	660
Mra22	*VIOLACEAE*	*Melicytus ramiflorus*	EF103472.1	472
Mra23	*VIOLACEAE*	*Melicytus ramiflorus*	EF103473.1	468
Mra24	*VIOLACEAE*	*Melicytus ramiflorus*	EF103471.1	465
Mra25	*VIOLACEAE*	*Melicytus ramiflorus*	EF103470.1	472
Mra26	*VIOLACEAE*	*Melicytus ramiflorus*	EF103474.1	480
Globa A	*VIOLACEAE*	*Gloeospermum blakeanum*	GQ438777.1	491

Tabla 2: Relación de las 312 secuencias de ciclótidos encontrados especificando su familia, subfamilia, número de acceso y número de aminoácidos

Ciclótido	Familia	Especie	N° Acceso	N° Nucleótidos
Cter B	*FABACEAE*	*Clitoria ternatea*	JF931998.1	584
Cter M	*FABACEAE*	*Clitoria ternatea*	JF501210.1	514
cliotide T8	*FABACEAE*	*Clitoria ternatea*	JF931994.1	515
cliotide T9	*FABACEAE*	*Clitoria ternatea*	JF931995.1	523
cliotide T2	*FABACEAE*	*Clitoria ternatea*	JF931989.1	494
cliotide T12	*FABACEAE*	*Clitoria ternatea*	JF931996.1	478
Panitide L1	*POACEAE*	*Steinchisma laxum*	KC182530.1	456
Panitide L2	*POACEAE*	*Steinchisma laxum*	KC182531.1	473
Panitide L4	*POACEAE*	*Steinchisma laxum*	KC182529.1	489
Panitide L6	*POACEAE*	*Steinchisma laxum*	KC182533.1	481
Phyb A	*SOLANACEAE*	*Petunia x hybrida*	JQ886398.1	629