UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS
FACULTAD DE MATEMÁTICA, FÍSICA Y COMPUTACIÓN

NEXOS ENTRE LA TAXONOMÍA EVOLUTIVA Y LA DISTRIBUCIÓN DE LAS FRECUENCIAS DE LOS AMINOÁCIDOS EN GENES Y PROTEÍNAS.

Tesis presentada en opción al título académico de Master en Matemática Aplicada.

Autor: Lic. María Milena Rodríguez Fernández

Tutor: Dr. Robersy Sánchez Rodríguez

Santa Clara
2008

Agradecimientos

A mi tutor Robersy Sánchez por su apoyo en todo momento

A mis padres y hermanas

Al Grupo de Bioinformática

Al Departamento de Matemática

A todos los que me han ayudado

Agradecimiento especial

A mi hija María Fernanda y a mi esposo por ser fuentes inspiradoras en cada paso por el camino de la vida

Resumen

En este trabajo se muestran evidencias acerca de la existencia de diferencias estadísticamente significativas entre varios grupos taxonómicos a través del uso del número de codones estimados que codifican para cada aminoácido (NEC_k) y de las probabilidades de aparición de estos en las proteínas (p_k). Estas variables fueron estimadas utilizando bases de datos construidas a partir del uso de codones en los genes y de secuencias de proteínas provenientes de diferentes organismos vivos. La aplicación de los métodos CHAID y análisis discriminate y la evaluación del desempeño de los mismos permitieron verificar que las diferencias detectadas entre los taxa están en correspondencia con la clasificación biológica. Además de esto, utilizando la distancia de Hellinger entre los vectores p_k se calcularon matrices de distancia a partir de las cuales se construyeron árboles filogenéticos que, no solo confirmaron relaciones filogenéticas entre los taxa que están en concordancia con la taxonomía evolutiva sino que, además, sugieren la existencia de, al menos, una gran extinción en algún momento de la historia evolutiva.

Abstract

Evidences about the existence of statistical significant differences between several taxonomic groups are shown by means of the number of codon used by each amino acid (NEC_k) and the appearance probabilities of these in proteins (p_k). These variables were estimated using databases which were made-up from the codon use in genes and protein sequences taken from different living organisms. The application of CHAID and discriminant methods and their performance evaluation allowed verify that the differences detected between the taxa are in correspondence with their biological classification. Besides this, distance matrices were calculated using the Hellinger distance between p_k vectors. From these matrices phylogenetic trees were built that, not only, confirmed the phylogenetic relationships between the taxa that are in agreement with the evolutionary taxonomy but rather, also, they suggest the existence of, at least, a great extinction during the evolutionary history.

TABLA DE CONTENIDOS

INTRODUCCIÓN

El desarrollo alcanzado por las Ciencias Biológicas ha permitido la acumulación de mucha información experimental disponible en grandes bases de datos. El análisis de los diferentes caracteres fenotípicos, como la morfología, la conducta, los cromosomas, la anatomía externa e interna, el desarrollo embrionario, el metabolismo, la variación genética y proteíca muestran que las especies presentan semejanzas homólogas en todos los niveles del fenótipo. Cuanto más próxima sean la especies, mayor será el grado de semejanza, y lo contrario también es cierto, cuanto más alejada estén menos semejanzas encontraremos.

Antecedentes y actualidad del tema.

La universalidad de la molécula portadora de la información genética hace que el DNA sea un carácter muy apropiado para el estudio comparativo y filogenético de las especies. Morfológicamente no es posible comparar una bacteria con un hombre, sin embargo si es posible establecer una comparación con moléculas de DNA de ambos organismos, ya que están formadas por el mismo lenguaje de bases. Con datos de secuencias podemos comparar cualesquier grupo de organismos, por distantes que sean. Los datos moleculares tienen otras propiedades adicionales que todas juntas los convierten en el carácter ideal de estudios filogenéticos. Muchos trabajos obtienen y analizan las secuencias de genes y proteínas de diferentes especies para resolver cuestiones todavía dudosas de relaciones entre organismos.

Formulación del problema

- ¿Constituyen los cambios en las frecuencias de aminoácidos huellas que permiten diferenciar las especies de organismos en correspondencia con su clasificación taxonómica?
- ¿Es posible mediante el análisis de distribución de probabilidades de la aparición de aminoácidos en las diferentes especies de organismos confirmar las relaciones filogenéticos entre los organismos y deducir nuevos aspectos del proceso de evolución molecular?

Hipótesis de trabajo

Teniendo en cuenta los elementos teóricos antes expuestos, así como las interrogantes existentes, se plantea la siguiente hipótesis general de investigación:

A pesar de que en el proceso de evolución molecular la mayoría de los genes codificantes para proteínas se originaron a partir de la combinación de regiones codificantes para dominios estructurales de las proteínas ancestros, es posible distinguir los taxa a partir de las diferencias estadísticamente detectables en el número estimado de codones que codifican para cada aminoácido.

Objetivo del trabajo

Este trabajo se propone como objetivo:

Detectar diferencias estadísticamente significativas en el número estimado de codones que codifican para cada aminoácido que estén en correspondencia con las clasificaciones taxonómicas existentes.

Tareas de investigación

Para el cumplimiento del objetivo y responder a las preguntas de investigación, demostrando la hipótesis anterior, fue necesario:

- Construir dos bases de datos con 9 grupos de organismos (archaea, bacterias, invertebrados, insectos, plantas, vertebrados que no son mamíferos, mamíferos que no son primates, primates y homo sapiens), ellas se componen una de cadenas de aminoácidos y otra de la frecuencia del uso de codones, ambas extraídas de Internet, la primera de **Direct public access to the National Library of Medicine's Medline Biomedical literature search engine through the NCBI. www.ncbi.nlm.nih.gov/entrez/ (PubMed)** y la segunda de **Codon Usage Database.**
- Calcular la frecuencia de aparición de cada aminoácido en la cadena representativa de los organismos**.**
- Determinar el número estimado de codones para subgrupos de organismos formando así vectores 20 dimensionales con los que se realizaron las pruebas estadísticas.

- Calcular las distancias genéticas entre los pares de poblaciones estudiadas donde cada distancia equivale al grado de divergencia proporcional entre las dos poblaciones, en la base de datos de aminoácidos, utilizando la distancia de Hellinger y la Entropía Relativa.
- Realizar el análisis estadístico aplicando la técnica del CHAID y Análisis de Discriminante.
- Evaluar la eficacia de los clasificadores a través de Curvas ROC y los parámetros que se obtienen de la matriz de confusión que nos permita sustentar la hipótesis de investigación.
- Construir los árboles filogenéticos referidos a la base de datos de aminoácidos, mostrando nuevos aspectos de las relaciones de evolución entre las especies.

Novedad Científica

La novedad científica del presente trabajo se resume en:

1. Se encuentran evidencias estadísticamente significativas acerca de la factibilidad del empleo del número estimado de codones que codifican para cada aminoácido en la clasificación taxonómica de los organismos vivos.
2. Se muestran la factibilidad del empleo de las estimaciones de las probabilidades de aparición de los aminoácidos en proteínas en la construcción de árboles filogenéticos y en la detección de posibles ancestros extintos durante grandes extinciones masivas.

Importancia teórica

En este trabajo se desarrolla un tratamiento alternativo de la información presente en las secuencias de genes y proteínas para su uso en el análisis taxonómico y filogenético. En particular, el tratamiento realizado permite prescindir de los posibles errores tautológicos derivados de los procesos de multialineación de secuencias de genes o de secuencias de proteínas, el cual es una etapa necesaria cuando se realizan los análisis mencionados partiendo de las secuencias biológicas.

Importancia práctica

La variabilidad de secuencias de genes y de proteínas de los múltiples organismos utilizadas en esta tesis, implica que la complejidad de los procedimientos matemático-

computacionales a realizar, para minimizar los errores tautológicos derivados de los multialinemientos de las secuencias biológicas, requiere de una logística computacional e intelectualmente multidisciplinaría muy costosa y poco frecuente en los grupos de trabajo de Bioinformática. Luego, el procedimiento que se propone en este trabajo puede resultar una alternativa muy útil.

Estructura del trabajo

La tesis ha sido estructurada de la siguiente forma: introducción, 4 capítulos, conclusiones, recomendaciones y anexos. En el capítulo 1 desarrollamos el marco teórico que le permita al lector un conocimiento general del tema abordado y la comprensión de los capítulos siguientes. En el capítulo 2 nos proponemos dra cumplimiento a la primera tarea de nuestro trabajo, explicando con detalles la conformación de estas bases de datos. El tercer capítulo está destinado a describir las pruebas estadísticas realizadas con el SPSS y un cuarto capítulo donde se expone lo relacionado con las relaciones evolutivas encontradas a partir del procesamiento de estas bases de datos. Finalmente aparecen las conclusiones del trabajo.

1. BASES BIOLÓGICAS Y MATEMÁTICAS

En este capítulo se realiza una descripción de las bases teóricas que conducen a las aplicaciones que tiene hoy en día el estudio molecular. Primeramente, se describe un sumario biológico con aquellos términos más usados, elementos importantes del código genético y aspectos biológicos de la evolución molecular. Posteriormente, se presentan una descripción de las herramientas estadísticas y bioinformáticas usadas en el trabajo.

1.1. Sumario Biológico Teórico

Algunos de los términos biológicos usados en el trabajo son lo que siguen:

Especie son agrupamientos de poblaciones naturales intercruzantes, con las mismas características, que ocupan una determinada área geográfica y están reproductivamente aisladas de otros grupos.

Taxón: (del griego *taxis* = arreglo, poner orden) Término aplicado a un grupo de organismos situado en una categoría de un nivel determinado en un esquema de clasificación taxonómica.

Taxonomía: (del griego *taxis* = arreglo, poner orden; *nomos* = ley): Método sistemático de clasificar plantas y animales. Clasificación de organismos basada en el grado de similitud, las agrupaciones representan relaciones evolutivas (filogenéticas).

Micro Taxonomía es la taxonomía que trata los organismos a nivel de especies y poblaciones.

Macro Taxonomía es la taxonomía que trata los organismos a nivel de las categorías superiores como género, familia, orden, etc.

El **proteoma** es el conjunto completo de proteínas que se expresan en el genoma. Algunos genes codifican para múltiples proteínas, el tamaño del proteoma es mayor que el número de genes. A veces el término se usa para describir el comportamiento de proteínas expresadas por una célula en un momento. Puede usarse para referirse al juego de proteínas codificadas por el genoma entero o en particular para cualquier célula o tejido.

El **genoma** es el juego completo de genes de un organismo. Se define por la sucesión de ADN completa, aunque en la práctica no puede ser posible identificar exactamente cada gen solamente en base a la sucesión que lo representa.

El **transcriptoma** es el juego completo de genes expresado bajo particulares condiciones. Se define como el juego de moléculas de ARN que están presentes, y puede referirse a un solo tipo de célula o a la unión más compleja de células o al organismo completo. Como algunos genes generan el mRNAs múltiple, es probable que el transcriptoma sea más grande que el número de genes definido directamente en el genoma.

Las proteínas pueden funcionar independientemente o como parte del multiprotein. Si se pudieran identificar todas las interacciones entre proteínas, podríamos definir el número total de dominios independientes de proteínas.

Ortólogos son las proteínas correspondientes en dos especies diferentes en sucesiones homólogas. Por lo general contamos que dos genes en organismos diferentes, proporcionan funciones correspondientes si sus secuencias son similares sobre el 80 % de la longitud, Figura 1.1.1. Según este criterio, el 20 % aproximadamente de los genes de mosca tiene ortólogos tanto en la levadura como en el gusano. Todo el reino eucariotes posiblemente requiere estos genes. La proporción aumenta al 30 % cuando la mosca y el gusano son comparados, representando la adición de las funciones que son comunes al reino eucariotes multicelular [1].

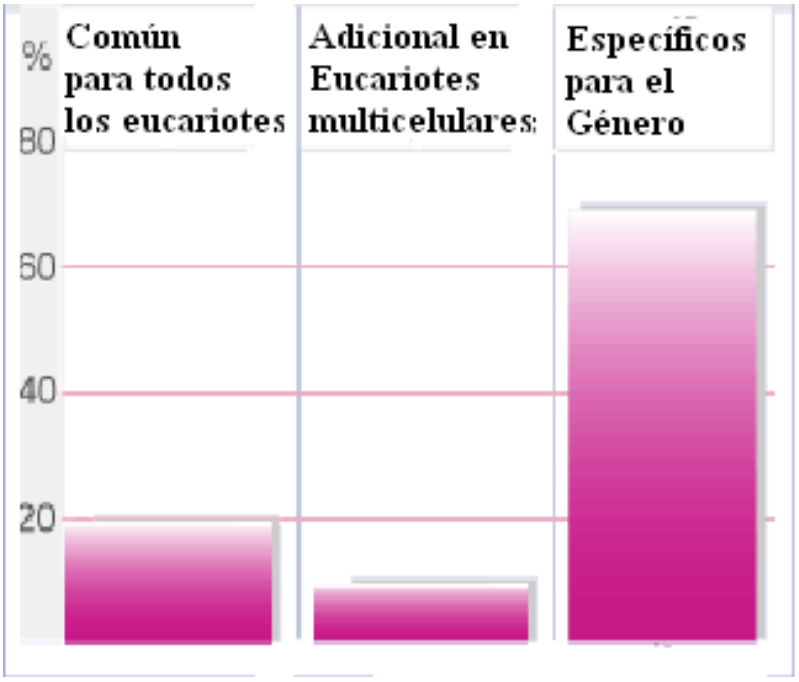


Figura 1.1.1. Las proteínas Ortólogos.

Los genes quehaceres domésticos (**gen Constitutivo**) son aquéllos (teóricamente) expresados en todas las células porque ellos proporcionan funciones básicas necesarias para el sustento de todos los organismos celulares.

El estudio de las secuencias genómicas puede ayudar a la comprensión de la función de las proteínas y los genes. Los estudios de proteína y evolución del gen involucran la comparación de sucesiones homólogas que tienen los orígenes comunes pero pueden o no tener una actividad común. Sucesiones que comparten un arbitrario nivel de similitud determinado por la alineación de emparejar las bases son homólogas. Ellos se heredan de un común antepasado que tenía estructura similar, aunque la estructura del antepasado puede ser difícil de determinar porque se ha modificado a través del descenso. **Homólogos** son la mayoría normalmente cualquier ortólogo, parálogos, o xenólogos.

Ortólogos son homólogos producidos por la especiación. Ellos representan genes derivados de un antepasado común que divergió, son asociados con la deuda de la divergencia de los organismos, tienden a tener función similar.

Parálogos son homólogos producidos por la duplicación del gen. Ellos representan los genes derivados de un gen hereditario común que se reprodujo dentro de un organismo y entonces como consecuencia divergido, tienden a tener las funciones diferentes.

Xenólogos son homólogos que son el resultado del traslado del gen horizontal entre dos organismos.

Arqueo bacterias (del griego *arkhaios* = antiguo; *bakterion* = bastón: grupo de procariotas de unos 3.500 millones de años de antigüedad, presentan una serie de características diferenciales que hicieron que Carl Woese profesor de la Universidad de Illinois, Urbana, U.S.A., proponga su separación del reino Moneras y la creación de uno nuevo: Archaea, propuesta que hoy es aceptada.

ARN ribosómico: Uno de los tres tipos de ARN, el ARNr es un componente estructural de los ribosomas. Son el "core" (parte principal) de los ribosomas y posiblemente la clave del mecanismo de traducción de las proteínas. Su estudio comparativo llevó a postulación de un Árbol Filogenético Universal.

Eubacterias (del griego *eu* = bueno, verdadero; *bakterion* = bastón): subgrupo del reino Monera que incluye a las bacterias verdaderas como *Escherichia coli*

Eucariotas (del griego *eu* = bueno, verdadero; *karyon* = núcleo, nuez): organismos caracterizados por poseer células con un núcleo verdadero rodeado por membrana. El registro arqueológico muestra su presencia en rocas de aproximadamente 1.200 a 1500 millones de años de antigüedad.

Filogenía (del griego *phylon* = raza, tribu):

1) el estudio de relaciones evolutivas en un grupo.

2) hipótesis evolutiva representada en un diagrama como un "árbol evolutivo".

3) estudio de la formación y la evolución de los organismos, con el objeto de establecer su parentesco.

Genes (del griego *genos* = nacimiento, raza; del latín *genus* = raza, origen): segmentos específicos de ADN que controlan las estructuras y funciones celulares; la unidad funcional de la herencia. Secuencia de bases de ADN que usualmente codifican para una secuencia polipeptídica de aminoácidos.

LUCA (del ingles, **L**ast **U**niversal **C**ellular **A**ncestor): antepasado común de las células modernas equivale a lo que es Lucy en el árbol evolutivo de *Homo sapiens*, es decir, no la primera célula sino una célula ya evolucionada, con todas las características de sus futuros descendientes: los actuales procariotas y eucariotas (ADN, Código genético, síntesis proteica etc.). Término propuesto en un coloquio de la Fundación Treille: http://www-archbac.u-psud.fr/Meetings/LesTreilles

Transferencia horizontal de genes: mecanismo por el cual se transmiten genes individuales, o grupos de ellos, de una especie a otra.

Secuencia conservada: Secuencia de base en una molécula de ADN (o de aminoácidos en una proteína) que ha permanecido prácticamente intacta a lo largo de la evolución.

Evolución paralela o convergente es la evolución de un carácter en dos o más especies, como la aptitud para volar, puede producirse de dos formas. El carácter puede aparecer en un ancestro común a ambas especies y transmitirse por herencia; en este caso se habla de homología. Los caracteres considerados podrían asimismo evolucionar de manera independiente en cada especie. En la evolución paralela se conserva el estado ancestral de las dos especies que comparten el carácter común; en la evolución convergente se modifica el estado ancestral. Por ejemplo la capacidad de volar se ha desarrollado de manera independiente en murciélagos, aves e insectos, además de en grupos ahora extinguidos y conocidos por sus fósiles, como los reptiles llamados pterosaurios. Todos estos animales han desarrollado alas por evolución convergente.

Una **extinción masiva** (también llamado evento a nivel de extinción o ELE por sus siglas en inglés) es un período de tiempo en el cual desaparece un número muy grande de

especies. Por el contrario, se estima que en períodos normales las especies desaparecen a un ritmo de entre dos y cinco familias biológicas de invertebrados marinos y vertebrados cada millón de años. Desde que la vida empezó en la Tierra se han detectado seis sucesos de extinción graves en el eón Fanerozoico.

1.2. El código genético y aspectos biológicos de la evolución molecular.

La Biología Teórica actual centra su atención en la investigación de las estructuras básicas de la vida. Una de estas estructuras básicas es el sistema bioquímico que hace posible el flujo de la información genética en los organismos vivos, el código genético. La relación entre las secuencias de ADN y las proteínas correspondientes es llamada código genético [1]. En este sistema se establecen las reglas mediante las cuales toda secuencia de nucleótidos del ADN, correspondiente a un gen, es transcripta en la secuencia de codones del ARNm y seguidamente es traducida en la secuencia de aminoácidos de la proteína correspondiente. Inicialmente se pensó que el código era universal −abarcando a todas las especies vivas− pero, posteriormente, fueron encontradas variaciones nucleares y mitocondriales [2] (para una revisión ver [3]). Sin embargo, estas variaciones son limitadas y corresponden esencialmente a reasignaciones de uno o varios codones a otros aminoácidos. Luego, el código genético puede ser considerado, con justicia, universal [4].

El código genético es la piedra angular del sistema de información genética. Consecuentemente, es de esperar que toda construcción teórica que intente explicar las relaciones cuantitativas y cualitativas existentes en el sistema de información genética tome como punto de partida el código genético. Lewin también define el código genético como la correspondencia entre los tripletes de bases en el ADN (o en el ARN) y los aminoácidos en las proteínas. En el código genético encontramos que los aminoácidos, excepto el Triptófano (W) y la Metionina (M), son codificados por más de un codón, por lo cual se dice que es un código degenerado. Las reglas mediante las cuales los aminoácidos fueron asignados a los tripletes de base que forman el código genético constituyen un enigma hasta el presente.El conjunto de tripletes de bases o codones que forman el código genético es una extensión del alfabeto de cuatro “letras” encontradas en la molécula del ADN.

Tabla 1.2.1. Tabla del código genético estándar [a, b].

Primera base del Codón	Segunda base del Codón: U			Segunda base del Codón: C			Segunda base del Codón: A			Segunda base del Codón: G			Tercera base del Codón
U	UUU	Phe	F	UCU	Ser	S	UAU	Tyr	Y	UGU	Cys	C	U
	UUC			UCC			UAC			UGC			C
	UUA	Leu	L	UCA			UAA	TER	-	UGA	TER	-	A
	UUG			UCG			UAG			UGG	Trp	W	G
C	CUU	Leu	L	CCU	Pro	P	CAU	His	H	CGU	Arg	R	U
	CUC			CCC			CAC			CGC			C
	CUA			CCA			CAA	Gln	Q	CGA			A
	CUG			CCG			CAG			CGG			G
A	AUU	Ile	I	ACU	Thr	T	AAU	Asn	N	AGU	Ser	S	U
	AUC			ACC			AAC			AGC			C
	AUA			ACA			AAA	Lys	K	AGA	Arg	R	A
	AUG	Met	M	ACG			AAG			AGG			G
G	GUU	Val	V	GCU	Ala	A	GAU	Asp	D	GGU	Gly	G	U
	GUC			GCC			GAC			GGC			C
	GUA			GCA			GAA	Glu	E	GGA			A
	GUG			GCG			GAG			GGG			G

[a] Los aminoácidos codificados por cada codón se representan con el símbolo de tres letras y el símbolo de una letra.

[b] El codón AUG es utilizado con mayor frecuencia como codón de inicio de la transcripción y codifica para el aminoácido Metionina (Met). Los codones UAA, UAG y UGA (TER) son marcadores del final de los genes.

Estas "letras" son las bases nitrogenadas del ADN: adenina, guanina, citosina y timina, las cuales son denotadas usualmente como A, G, C y T respectivamente (en la molécula del ARN la base T es cambiada por el uracilo, U). En la doble hélice formada por el ADN, la base G es complementaria de la base C y la base A es complementaria de la base T. Estas bases están apareadas en la doble hélice de acuerdo con la siguiente regla: G≡C, A≡T, donde '–' simboliza un puente de hidrógeno.

El código genético estándar (Tabla 1.2.1) puede ser considerado, con toda justicia, universal [4], pues solo existen algunas variaciones en mitocondrias, bacterias y algunos eucariotes unicelulares (para una revisión ver [3]). Sin embargo, estas variaciones son limitadas y corresponden esencialmente a reasignaciones de uno o varios codones a otros

aminoácidos. Los códigos genéticos conocidos han sido usualmente representados en tablas de cuatro entradas donde los codones están localizados atendiendo a la segunda base. Estas tablas pueden encontrarse en la página web (del NCBI): http://130.14.29.110/Taxonomy/Utils/wprintgc.cgi?mode=c.

En la tabla del código genético estándar tres entradas corresponden a los cambios de bases en los codones, realizados de acuerdo a diferentes criterios. Como resultado, los aminoácidos hidrofóbicos e hidrofílicos quedan localizados en diferentes columnas. De la observación del código genético se destaca que la degeneración del código implica solamente a la tercera posición del codón en la mayoría de los casos (son excepciones la Arginina (R), la Leucina (L) y la Serina (S) (Tabla 1.2.1). De esta forma resulta que las dos primeras bases de cada codón son las determinantes principales de su especificidad. La posición tercera, esto es, el nucleótido situado en el extremo 3' del codón tiene menor importancia y no encaja con tanta precisión, *está suelto y tiende a* "*vacilar*" según expresiones de F. H. C. Crick [2]. De manera que en la tabla del código estándar localizamos una partición natural en cuatro grupos de aminoácidos atendiendo a la base encontrada en la segunda posición: los aminoácidos cuyos codones poseen en la segunda posición la base U, los que poseen A, los que poseen G y los que poseen C (Tabla 1.2.1). Esta partición resalta una diferencia en las propiedades fisicoquímicas de los aminoácidos; por ejemplo, los aminoácidos que tienen U en la segunda posición de sus codones son hidrofóbicos: {I, L, M, F}[1], mientras que los aminoácidos que tienen A en la segunda posición son hidrofílicos (también conocidos como aminoácidos polares): {D, E, H, N, K, Q, Y} [2]. Tales observaciones llevaron a Epstein señalar que los aminoácidos afines deben tener alguna relación extendida entre sus codones [5, 6]. Las regularidades observadas en el código genético –ampliamente discutidas en [2]– sugirieron desde su descubrimiento que la asignación de los aminoácidos a los codones no debió ocurrir al azar [2, 7]. La tendencia a representar aminoácidos similares por codones similares minimiza los efectos de las mutaciones. Este hecho incrementa la probabilidad de que un simple cambio de base no resulte en la sustitución de un aminoácido por otro o, al menos, involucre aminoácidos con propiedades fisicoquímicas similares [2].

[1] Aunque no es indispensable para la comprensión de texto, si el lector está interesado, el nombre del aminoácido correspondiente a cada símbolo lo puede encontrar en la sección 2.2.

Por otra parte algunos autores han planteado que el código genético está optimizado y fijado [7, 8]; aunque autores, como Woese y Gillis y colaboradores, han sugerido que el código genético pudo ser optimizado para limitar los errores en los procesos de transcripción y de traducción [4, 9]. En realidad, parece que el código genético ha evolucionado en la dirección de minimizar las consecuencias de los errores producidos durante la transcripción y la traducción [10]. Un código genético óptimo se refiere a una asignación óptima de los aminoácidos a los codones de manera tal que los efectos negativos causados por los eventos mutacionales durante el proceso de evolución molecular sean minimizados [23,24].

La importancia de la posición de las bases es sugerida por las frecuencias de errores encontradas en los codones. En otras palabras, los errores −mutaciones fijadas en la población de genes− en la tercera base del codón son más frecuentes que en la primera y estos a su vez son más frecuentes que los errores en la segunda base [9, 11-13]. Estas posiciones son, sin embargo, más conservativas con respecto a los cambios en la polaridad de los aminoácidos codificados [14]. Como consecuencia, los efectos de las mutaciones están reducidos en los genes y las mutaciones fijadas en la población decrecen desde la tercera base a la segunda.

En los diferentes organismos (especies) existen diferencias en cuanto al uso que se hace de cada codón [15]. Se ha determinado que existe un uso preferencial de algunos codones sinónimos sobre otros, de manera que algunos codones son más frecuentemente usados que otros (ver por ejemplo. http://www.kazusa.or.jp/codon) y cada especie tiene sus codones “preferidos” o codones más frecuentemente usados. Esto significa que muestran un sesgo en el uso de los codones sinónimos. El uso de los codones no es al azar y puede estar asociado a varios factores tales como el nivel de expresión genética [16], la longitud del gen [17] y la estructura secundaria de las proteínas [17- 21]. Y aún más, para la mayoría de los aminoácidos en todas las especies vivas existe una asociación altamente significativa con la función del gen correspondiente, indicando que, en general, el uso de codones al nivel de aminoácidos individuales está estrechamente coordinado con la función del gen [22]. Esto nos sugiere que para los codones existen diferencias cuantitativas en valores que son expresados en las secuencias de codones de los genes. Estas diferencias cuantitativas precisamente nos posibilitan una descripción formal,

mediante modelos matemáticos, de las relaciones existentes entre los codones y entre los genes.

1.3. Herramientas estadísticas y bioinformáticas

El cálculo de probabilidades suministra las reglas apropiadas para cuantificar la incertidumbre y constituye la base para la estadística inductiva o inferencial. Las medidas no paramétricas de divergencia entre distribuciones de probabilidad se definen como expresiones funcionales, que miden el grado de discrepancia entre dos distribuciones cualesquiera, no necesariamente pertenecientes a una misma familia paramétrica. Después de los trabajos pioneros de Pearson (prueba ji-cuadrado) y Hellinger (la famosa distancia de Hellinger, publicada en 1909), medida de distancia definida sobre el espacio de las distribuciones de probabilidad, otros autores han estudiado divergencias (Shannon, Kullbach y Leibler, Renyi, etc). La divergencia aplicada a distribuciones de probabilidad serían introducidas por Csiszar (1963, 1967, 1972, 1975), estudiadas en diferentes versiones por Matusita (1955, 1964), Havrda y Charvat (1967), Vajda (1972) y generalizadas por Burbea y Rao (1982*).* Las divergencias tienen aplicaciones en inferencia estadística y en procesos estocásticos.

Para desarrollar nuestra investigación como herramientas estadísticas se usaron del SPSS el análisis CHAID y el Discriminante, para evaluar el desempeño de estos clasificadores se realizaron las curvas ROC y se calcularon los parámetros a partir de la matriz de confusión. Como herramienta Bioinformática se uso el MEGA4. A continuación describimos algunos aspectos técnicos de estas herrmientas.

1.3.1. CHAID, Chi-squared Automatic Interaction Detector

El método detector de interacciones basado en chi-cuadrado (CHAID) surge como una técnica de segmentación [33]. Su propósito es segmentar o dividir una población en dos o más grupos en las categorías del mejor predictor de una variable dependiente. El algoritmo se basa en la prueba chi-cuadrado para seleccionar la mejor división en cada paso, la división se realiza hasta que no haya más variables predictoras significativas o hasta que se satisfaga algún otro criterio de parada, relacionado por ejemplo con el número mínimo de casos en un nodo para analizar su divisibilidad.

En un estudio real existen frecuentemente múltiples variables (predictivas o independientes) que pueden tener asociación con una variable dependiente y además efectos de interacción entre ellas sobre dicha variable dependiente. La presentación de muchas tablas de contingencia, no siempre refleja las asociaciones esenciales, y usualmente se convierte en un listado inútil de tablas que desinforman en lugar de orientar, aun cuando se utilicen estadísticos (como la *V* de Cramer) para ordenar la fortaleza de las asociaciones. Un estudio multivariado trata de enfocar el efecto posible de todas las variables conjuntamente incluyendo sus posibles correlaciones; pero puede ser particularmente interesante, si considera además la posibilidad de la interacción entre las variables predictivas sobre la variable dependiente. Cuando el número de variables crece, el conjunto de las posibles interacciones crece en demasía, resulta prácticamente imposible analizarlas todas y por ello adquiere especial interés una técnica de detección automática de interacciones fundamentales. CHAID es exactamente eso, es útil en todos aquellos problemas en que se quiera subdividir una población a partir de una variable dependiente, y posibles variables predictivas que cambien los valores de la variable dependiente en cada una de las subpoblaciones o segmentos. La técnica de CHAID es capaz de segmentar la población en grupos de acuerdo con determinados valores de las variables y sus interacciones que distinguen de forma óptima, diferencias esenciales en el comportamiento de la variable dependiente (CHAID 1994).

Un análisis de CHAID automático comienza dividiendo la población total en dos o más subgrupos distintos basado en las categorías del mejor predictor de la variable dependiente (en principio por el estadígrafo chi-cuadrado de Pearson) [27]. Divide cada uno de estos subgrupos en pequeños sub-subgrupos y así sucesivamente. CHAID visualiza los resultados de la segmentación en forma de un diagrama tipo árbol cuyas ramas (nodos) corresponden a los grupos (subgrupos conformados en cada nivel). Entiéndase en este caso que está seleccionando sucesivamente las variables más significativamente asociadas con la clase y las variables que deben ser fuentes de estratificaciones sucesivas.

1.3.2. Analisis Discriminante.

Las técnicas de comparación Multivariada que se basan en particular en la construcción de una función de clasificación –conocida como análisis discriminante – han sido

desarrolladas recientemente comparadas con otras técnicas. Las primeras ideas surgen en la cuarta década del siglo XX, relacionadas precisamente con investigaciones biológicas y antropométricas, y desarrolladas fundamentalmente por Mahalonobis (1930) y Fischer (1936).

Son las técnicas de comparación Multivariada más ricas porque permiten la distinción general de los grupos, la determinación del orden de importancia de las variables discriminantes o distintivas y la precisión de una variable aleatoria discreta (Grupo) respecto a m variables en principio continuas o al menos ordinales: $X_1 \quad X_2 \quad \cdots \quad X_m$

Para determinar el orden de importancia de las variables $(X_1, X_2, \cdots, X_m)$ en la clasificación, se puede utilizar el coeficiente de correlación de estas variables con la función discriminante y tener una medida de las posibilidades de error.

La interpretación de la no presencia de una variable en la ecuación no puede ser interpretada como la independencia del proceso de clasificación respecto a esta variable pues de hecho en la ecuación puede haber otras variables que se correlacionan fuertemente con ésta. En definitiva la importancia absoluta de una variable la sigue brindando la significación del coeficiente de correlación de esa variable con la función, esté o no ella en la ecuación. Usualmente se exige que el por ciento de casos bien clasificados del total de la muestra no sea inferior a un 75% para que el criterio de clasificación sea considerado bastante bueno; pero este porcentaje “mínimo” puede variar sobre todo en el sentido de ser más exigente, en dependencia de los requisitos y características de la investigación.

La lambda de Wilks es otro estadístico que permite evaluar la hipótesis de que dos o más grupos provienen de poblaciones con las mismas medias para un conjunto de variables. El valor de esta lambda siempre está entre 0 y 1. Grandes valores de lambda indican que los grupos no parecen ser diferentes (en el caso de lambda igual a 1 los grupos fueran el mismo). Valores de lambda pequeños indican diferencias entre las medias de grupos. Precisamente por esto en cada paso del análisis discriminante se introduce la variable que más contribuye a la reducción de lambda entre los grupos. El estadístico lambda a veces se refiere en la literatura como estadístico U del análisis multivariado y se considera uno de los mejores criterios de comparación Multivariada y poco sensible a hipótesis de normalidad.

Existen varios métodos de análisis discriminante que pueden conducir a diferentes funciones de clasificación. En general estos métodos parten de hipótesis de normalidad conjunta de la variable vectorial $\vec{X}$ entre los grupos; pero en última instancia y sobre todo, en la normalidad de la variable que define la función discriminante:

$$F = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_m X_m$$

Esto permite que podamos utilizar variables $X_j \quad i = 1, \vec{m}$ que no tienen necesariamente una distribución normal conjunta, de hecho podemos utilizar incluso variables ordinales siempre y cuando la función discriminante resultante cumpla las condiciones de normalidad. Si se desea utilizar una variable nominal con *k* valores posibles, es conveniente sustituirla por *k* – 1 variables con valores (-1, 0, 1) como se hace en la regresión lineal múltiple.

La validez del análisis discriminante es menos sensible a la violación de la hipótesis de homogeneidad de covarianza si los volúmenes de las muestras son iguales. Se recomienda por ello utilizar diseños equilibrados.

1.3.3. El desempeño de los clasificadores usados. Matrices de confusión y las curvas ROC (Curva característica de operación del receptor)

El desempeño de un clasificador y sus diferentes alternativas de uso son validadas siguiendo los criterios clásicos de evaluación, en el trabajo se usan los parámetros de las matrices de confusión y las curvas ROC.

Las matrices de confusión contienen información acerca de los valores reales y las clasificaciones predichas hechas por cualquier sistema de clasificación. El desempeño de un sistema es usualmente evaluado usando los datos en dicha matriz.

		Clase verdadera	
		Pos	*Neg*
Clase Predicha	*pos*	*TP*	*FP*
	neg	*FN*	*TN*
Total columna		*P*	*N*

Figura1.3.3.1. Matriz de confusión.

En la Figura1.3.3.1 se muestra la matriz de confusión de un problema para dos clases, donde *Pos/pos* es la clase positiva y *Neg/neg* la clase negativa; *TP* y *TN* son los elementos bien clasificados de la clase positiva y negativa respectivamente. *FP* y *FN* son los elementos negativos y positivos mal clasificados respectivamente.

Han sido definidos varios términos estándar para medir el desempeño de un clasificador [26], de uso general en cualquier rama donde se apliquen sistemas de clasificación:

La **Exactitud** (Ac, del inglés Accuracy) es la proporción del número total de predicciones que fueron correctas: $Exactitud = \frac{TP+TN}{P+N}$

La **Razón de Verdaderos Positivos** (TP, del inglés True Positive Rate), es la proporción de casos positivos que fueron correctamente identificados:

$tp \quad rate = \frac{TP}{P} = recall = sensitivity$

La **Razón de Verdaderos Negativos** (TN, del inglés True Negative Rate) es la proporción de casos negativos que han sido correctamente clasificados:

$tn \quad rate = \frac{TN}{N} = specificity$

Finalmente, la **Precisión** (P, en inglés, también Precisión) es la proporción de casos predichos positivos que fueron correctos: $precisión = \frac{TP}{TP+FP}$

La **Razón de Falsos Negativos** (FN, del inglés False Negative Rate es la proporción de casos positivos que fueron incorrectamente clasificados como negativos: $fn \quad rate = \frac{FN}{P}$

La **Razón de Falsos Positivos** (FP, del inglés False Positive Rate) es la proporción de casos negativos que han sido incorrectamente clasificados como positivos: $fp \quad rate = \frac{FP}{N}$

Cuando el problema de clasificación abarca más de 2 clases, digamos tres clases, hay una TP rate para cada clase.

Otra forma de evaluar el rendimiento de un clasificador es por las curvas ROC (*Receiver Operator Characteristic*, *Curva característica de operación del receptor*) (Fawcett 2004) [25]. En esta curva se representa el valor de razón de TP vs la razón de FP, mediante la variación del umbral de decisión. Se denomina umbral de decisión a aquel que decide si una instancia *x*, a partir del vector de salida del clasificador, pertenece o no a cada una de las clases. Usualmente, en el caso de dos clases se toma como umbral por defecto 0.5; pero esto no es siempre lo más conveniente. Se usa el área bajo esta curva, denominada AUC (*Área Under the Curve, área bajo la curva ROC*) como un indicador de la calidad del clasificador. En tanto dicha área esté más cercana a 1, el comportamiento del clasificador

está más cercano al clasificador perfecto (aquel que lograría 100% de TP con un 0% de FP).

Una curva ROC es un gráfico con la Razón de Falsos Positivos (FP=1-Sp) en el eje X y la Razón de Verdaderos Positivos (TPrate) en el eje Y. Las curvas quedan en el cuadrado [0,1] x [0,1]. El vértice superior izquierdo de este cuadrado: (0,1) representa al clasificador perfecto porque clasifica todos los casos positivos y todos los casos negativos correctamente pues FPrate=0 y TPrate=1. El vértice inferior izquierdo (0,0) representa un clasificador que predice todos los casos como negativos, mientras que el vértice superior derecho (1,1) corresponde a un clasificador que predice todos los casos como positivos. El punto (1,0) es un clasificador pésimo o estúpido que resulta incorrecto en todas las clasificaciones.

Una curva (o un punto) ROC es independiente de la distribución de las clases o el costo de los errores, es decir, no depende de que en la base de aprendizaje haya más casos negativos que positivos o viceversa.

Una curva ROC resume toda la información contenida en la matriz de confusión ya que FNrate es el complemento de TPrate y TNrate es el complemento de FPrate. Las curvas ROC constituyen una herramienta visual para examinar el equilibrio entre la habilidad de un clasificador para identificar correctamente los casos positivos y el número de casos negativos que están incorrectamente clasificados.

El área bajo la curva ROC puede ser usada como una medida de la exactitud en muchas aplicaciones. Si se comparan dos clasificadores, a través de sendas curvas ROC podemos decidir en general que la de mayor área bajo ella identifica al mejor clasificador.

Cuando el problema de clasificación abarca más de 2 clases, digamos tres clases, habrá que hacer una curva ROC para cada clase y se tendrá un área bajo cada una de las curvas.

1.3.4. MEGA: “Molecular Evolutionary Genetics Analysis”

MEGA es un instrumento integrado para conducir la alineación de secuencia automática y manual, deduciendo filogenéticamente árboles, extrayendo de bases de datos de web, estimando las tarifas de evolución molecular y probando hipótesis evolutivas [32].

Las relaciones filogenéticas de genes u organismos normalmente se presentan en árboles formados con una raíz que se llama un árbol arraigado. También es posible dibujar un árbol sin una raíz. El modelo de la bifurcación del árbol se llama una topología.

Hay numerosos métodos para construir los árboles filogenéticos de datos moleculares (Nei y Kumar 2000). Ellos pueden ser clasificados en los métodos de Distancia, métodos de parsimonia y métodos de Probabilidad.

UPGMA es un método que asume que la proporción de nucleotide o substitución del aminoácido es el mismo para todos los linajes evolutivos. Un aspecto interesante de este método es que él produce un árbol que imita un árbol de la especie. El MEGA4 brinda la posibildad al usuario de introducir su propia matriz de distancia para construir los árboles filogenéticos. Esta posibilidad fue aprovechada en nuestro trabajo.

2. CONSTRUCCIÓN DE LAS BASES DE DATOS Y PREPARACIÓN DE LAS MISMAS

La comparación de la sucesión del genoma humana con sucesiones encontradas en otras especies de organismos vivos es revelador del proceso de evolución, en nuestro trabajo se construyen dos bases de datos con 9 grupos de organismos(archaea, bacterias, invertebrados, insectos, plantas, vertebrados que no son mamíferos, mamíferos que no son primates, primates y homo sapiens), ellas se componen una de cadenas de aminoácidos ver Anexo2 y otra de la frecuencia del uso de codones ver Anexo 3, ambas extraídas de Internet, la primera de Direct public access to the National Library of Medicine's Medline Biomedical literature search engine through the NCBI. www.ncbi.nlm.nih.gov/entrez (PubMed) y la segunda de Codon Usage Database. Los grandes bancos de datos existentes en el mundo, dentro de los que se encuentran los usados por nosotros, se caracterizan por reunir las proteínas con gran variedad, dentro de las que se encuentran aquellas que podrían falsear nuestra información por su carácter de proteína conservadas dentro del proceso evolutivo de las especies, por lo que se realizó un minucioso trabajo de selección de las proteínas representativas en cada especie en cuestión. Además de contar con una representatividad de organismos y de proteínas en cada grupo, consideramos necesario explicar que los resultados obtenidos en la investigación muestran en determinados momentos aquellos datos presentes en la literatura, [1] como ejemplo en la Figura 2.1, donde se analizan los genes según su distribución en la naturaleza. Empezando con los más representativos, 21% de genes son comunes a eucariotas y procariotas. Éstos tienden a codificar para proteínas que son esenciales para todos los organismos vivientes - el metabolismo típicamente básico, repetición, trascripción, y traducción.

Moviéndonos en el sentido de las agujas del reloj, el 33% de genes se presentan generalmente en los organismos eucariotes. Éstos tienden a codificar para las proteínas involucradas en funciones que son generales a las células eucariotas pero no a las bacterias - por ejemplo, ellos pueden tener relación con especificar organelas o componentes del citoesqueleto. Otro 24% de genes sin especificar los vertebrados que incluyen son necesarios para el multicelularismo y para el desarrollo de diferentes tipos de tejidos. Y el 22% de los genes son únicos de los vertebrados mamíferos.

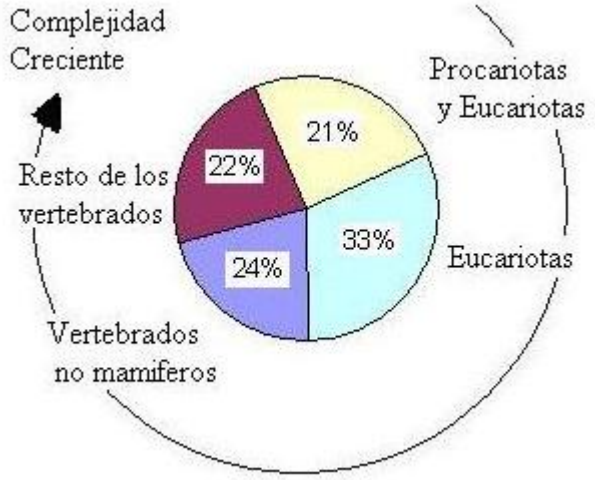


Figura 2.1. Distribución de los genes en la naturaleza según las funciones necesarias para la vida.

Éstos principalmente codifican para las proteínas de los sistemas inmune y nervioso; ellos codifican para muy pocas enzimas, relacionado con la idea que las enzimas tienen orígenes antiguos, y que las funciones metabólicas se originaron temprano en el proceso de evolución. Observamos, por consiguiente, que la progresión de las bacterias a los vertebrados requiere la suma de grupos de genes que representan las nuevas funciones necesarias en cada fase.

Una manera de definir las proteínas normalmente necesitadas es identificar las proteínas presentes en todos los proteomas [1]. Comparando el proteoma humano en más detalle con los proteomas de otros organismos, 46% del proteoma de levadura, 43% del proteoma del gusano, y 61% del proteoma de la mosca están presentes en el proteoma humano. Un grupo importante de aproximadamente 1300 de las proteínas están presentes en los cuatro proteomas. Las proteínas comunes son básicas, aquellas requeridas para las funciones esenciales lo cual queda resumido en Figura 2.2. Las funciones principales se representan por la trascripción y la traducción (35%), metabolismo (22%), transporte (12%), repetición de ADN y la modificación (10%), proteína de plegado y degradación (8%), y el resto representan otros procesos celulares.

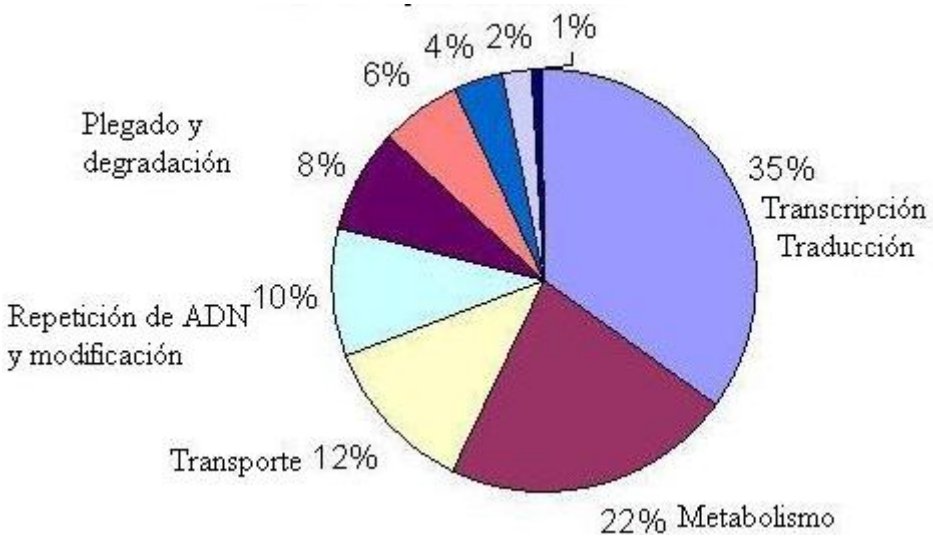


Figura 2.2 Distribución de los genes según los procesos celulares que realizan.

Uno de los rasgos llamativos del proteoma humano es que tiene muchas nuevas proteínas comparándolo con otros organismos eucariotes, pero tiene relativamente pocos nuevos dominios de la proteína. La mayoría de los dominios de las proteínas parecen ser comunes al reino animal. Hay sin embargo, muchas nuevas arquitecturas de la proteína, definidas como las nuevas combinaciones de dominios.

Además de los genes funcionales, hay también copias de genes que se han convertido en genes no funcionales (identificados como tal por las interrupciones en las sucesiones de proteína codificadas). Éstos se llaman pseudos genes (vea *Molecular Biology 1.4.6 Pseudogenes are dead ends of evolution*). El número de pseudos genes puede ser grande. En el ratón y en el genoma humano, el número de pseudo genes es aproximadamente el 10% del número de genes potencialmente activos.

Teniendo en cuenta las razones expuestas anteriormente, la selección de las bases de datos fue primordial para el logro de nuestros objetivos.

2.1. Construcción de las bases de datos.

La base de datos de proteínas para el entrenamiento esta formada por los nueve grupos nombrados anteriormente cada uno con aproximadamente 1000 cadenas y una variedad en cuanto a organismos y tipos de proteínas ver Tabla 2.1.1, esta base es nombrada en el trabajo como base curada por la selección minuciosa tanto de los organismos como de los

tipos de proteínas que la forman. Luego se confeccionó una base de datos para validar resultados, esta la nombramos base no curada la que esta formada por los mismos nueve taxa nombrados, pero con alrededor de 500 cadenas de aminoácidos y sin hacer ningún tipo de selección en lo que se refiere a proteínas que la forman. Para los análisis realizados en las taxa: archaea-bacteria, archaea-bacteria-eucariotes, vertebrados-invertebrados, vertebrados no mamiferos-mamiferos y homo sapiens-primates, se uso el 70% de la base que llamamos extendida que está constituida por la unión de la base curada más la no curada. Además en el caso de la taxa archaea-bacteria-eucariotes se uso la base extendida aumentada en número de cadenas pero sin tener en cuenta ningún tipo de selección de las proteínas ni de los organismos que la forman. En el caso de las taxa vertebrados-invertebrados y vertebrados no mamiferos-mamiferos se reorganizaron todos aquellos organismos que pertenecen a estos grupos aumentando asi el número de secuencias.

Tabla 2.1.1. Bases de datos.

Proteinas							**Uso de Codones**	
	Base curada		**Base no curada**		**Base extendida**		**Base curada**	
Grupos de org.	**No. de Sec.**	**50 subgrp.**	**No. de Sec.**	**20 subgrp.**	**Subgrps/ No. de Sec.**	**No. de Sec.**	**No. de Sec.**	**50 subgrp.**
Archaea	**1566**	**31**	**736**	**36**	**200/139**	**27844**	**1317**	**26**
Bacterias	**1334**	**26**	**449**	**22**	**200/28**	**5667**	**830**	**26**
Eucariotes					**200/124**	**24927**		
Invertebrados	**1221**	**24**	**768**	**38**	**100/37**	**3742**	**2187**	**43**
Insectos	**1010**	**20**	**743**	**37**			**979**	**19**
Plantas	**1762**	**35**	**488**	**24**			**2114**	**42**
Vertebrados					**100/93**	**9387**		
Vertebrados no Mamiferos	**1498**	**29**	**440**	**21**	**100/19**	**1938**	**1465**	**29**
Mamiferos					**100/74**	**7450**		
Mamiferos no Primates	**1593**	**31**	**519**	**25**			**2036**	**40**
Primates	**1473**	**29**	**394**	**19**	**70/26**	**1867**	**1831**	**36**
Homo Sapiens	**897**	**28**	**1162**	**55**	**70/20**	**2059**	**1821**	**36**

En la base de datos referida al uso de codones contamos al igual que la anterior con el mismo número de taxa cada una con alrededor de 1000 cadenas y la variedad en cuanto a tipo proteínas y especies que la forman, ver Tabla 2.1.1. Los análisis se realizaron solo con la base curada, no se formaron bases externas pues ello hubiera requerido mayor tiempo y otros objetivos no trazados en este trabajo**.**

2.2. Cálculo de los vectores NEC_k a partir de las bases de secuencias.

Los sistemas vivos, jerarquizados son altamente complejos ya desde el inicio de la vida misma [35]. Una bacteria tiene un sistema génico complejísimo muy similar al de cualquier Metazoo [28]. Las formas más elementales de estos comparten con los Metazoos más evolucionados, como los mamíferos, idénticas porciones de sus genomas. Estos hechos conducen directamente al planteamiento del problema que da lugar al trabajo de tesis, el cual acarrea definir el concepto de "número estimado de codones" (*NEC*). El concepto de *NEC* es derivado de la degeneración del código genético estándar (CGS) y de la existencia de un uso diferenciado de codones para cada especie (ver sección 1.1). Si se supone que el proceso de síntesis de proteínas ha sido optimizado y adaptado a las variaciones ambientales durante el proceso de evolución molecular que dio lugar a la especiación entonces, se debe esperar que, mientras mayor sea la frecuencia observada de un aminoácido f_{aa} en los genomas de los organismos vivos, mayor será, en general, su representación en la tabla del CGS. El NEC_k que codifican para el aminoácido k puede definirse como:

$$NEC_k = \frac{f_{aa}^{k}}{\sum_{i=1}^{20} f_{aa}^{i}} 61 \qquad (2.2.1)$$

donde $k = 1,\ldots, 20$ y el número 61 hace referencia al número total de codones en la tabla del CGS que codifican para los aminoácidos. Como se muestra en la Tabla 2.2.1 existe una correlación positiva entre el *NEC* y las frecuencias f_{aa} en las proteínas y los genomas de Archaeas, Bacterias y Eucariotes. Sin embargo, las frecuencias f_{aa} deben de estar afectadas por el uso de codones (ver sección 1.1), de manera que, como se aprecia en la Tabla 2.2.1, para cada aminoácido el *NEC* difiere en alguna medida del número de codones que codifican para dicho aminoácido en la tabla del CGS (ver Tabla 1.2.1 y Tabla 2.2.1).

Tabla 2.2.1. Correlación entre el *NEC* y las frecuencias f_{aa} en las proteínas y los genomas de Archaeas, Bacterias y Eucariotes.[a]

Aminoácido	No. Codones[b]	Archaeas %	Bacterias %	Eucariotes %	Todos	Frec. Aa [c]
Ala	4	4.789	4.929	3.953	4.758	4.697
Arg	6	3.611	3.044	3.196	3.190	3.111
Asp	2	3.337	3.087	3.239	3.166	3.172
Asn	2	2.074	2.824	2.904	2.666	2.623
Cys	2	0.543	0.610	1.135	0.671	1.220
Glu	2	4.752	3.874	4.050	4.099	3.782
Gln	2	1.159	2.373	2.611	2.105	2.501
Gly	4	4.569	4.087	3.587	4.130	4.514
His	2	1.037	1.263	1.470	1.238	1.403
Ile	3	4.630	4.301	3.343	4.240	3.233
Leu	6	5.887	6.417	5.704	6.192	5.551
Lys	2	3.684	3.922	3.843	3.855	3.599
Met	1	1.519	1.336	1.421	1.391	1.464
Phe	2	2.440	2.788	2.562	2.678	2.440
Pro	4	2.702	2.434	3.142	2.599	3.111
Ser	6	3.617	3.770	5.185	3.941	4.209
Thr	4	2.910	3.142	3.398	3.123	3.599
Trp	1	0.628	0.671	0.689	0.665	0.854
Tyr	2	2.245	1.970	1.848	2.013	1.952
Val	4	4.862	4.191	3.715	4.276	4.026
Coef. Corr. Pearson [d]		0.634	0.643	0.743	0.666	0.735

[a] Frecuencias de aminoácidos en 8 genomas de archaeas, 22 genomas de bacterias y 5 genomas de eucariotes [29].
[b] Número de codones que codifican para cada aminoácido en la tabla del CGS (ver Tabla 1.2.1).
[c] Frecuencias de aminoácidos en proteínas [30].
[d] Todas las correlaciones son altamentente significativas ($p < 0.01$).

Luego, el *NEC* constituye una variable que expresa la divergencia existente entre el CGS y el número efectivo funcional de codones. Por ejemplo, en la Tabla 2.2.1 el Ácido Glutámico (Glu) en Eucariotes posee un *NEC*=4.050, sin embargo en el CGS solo dos codones codifican para este aminácido. Esto no significa que en los organismos eucariotes existen más de dos codones que codifican para el Ácido Glutámico (pues solo hay dos), sino sugiere que, funcionalmente durante la síntesis de proteínas, se garantiza el material necesario (tRNA, enzimas involucradas, etc) para producir un efecto en la eficiencia del proceso de síntesis equivalente al que tendría la existencia de más de dos codones codificantes para dicho aminoácido.

Estos análisis nos sugieren utilizar la variable *NEC* tal y como se plantea en la hipótesis de investigación. Luego, los vectores NEC_k (20-dimensionales) se calcularon (Anexo 8) a partir de las secuencias de proteínas y del uso de codones que conforman las bases de datos descritas en las secciones anteriores. Con este propósito cada base de secuencias de proteínas fue particionada en subconjuntos de secuencias en correspondencia con su tamaño.

$$f_{aa}^{i} = \sum_{k=1}^{20} f\left(A_i, B_k\right) \tag{2.2.2}$$

Como consecuencia a cada taxa le corresponde un conjunto de vectores NEC_k los cuales fueron utilizados en las pruebas estadísticas que se realizaron para verificar la hipótesis de investigación.

Cuando se parte del uso de codones cada subconjunto de la partición está formado por vectores 64-dimensionales, cada uno de los cuales contiene las frecuencias de uso de los 64 codones del gen que representa. Si n_i ($\sum_{i=1}^{20} n_i = 61$) denota el número de codones que codifican para el aminoácido i (i =1,..,20), k denota el k-ésimo vector que contiene las frecuencias f_{jk}^{i} (j = 1,.., n_i) de uso de los 64 codones presentes en el k-ésimo gene, entonces la frecuencia observada f_{aa}^{i} del aminoácido i en un subconjunto conformado por m genes se estimó como:

$$f_{aa}^{i} = \sum_{k=1}^{m} \sum_{j=1}^{n_i} f_{jk}^{i} \tag{2.2.3}$$

Como consecuencia, a cada taxa le corresponde un conjunto de vectores NEC_k estimados por la expresión (2.2.1), los cuales fueron utilizados en las pruebas estadísticas que se realizaron para verificar la hipótesis de investigación.

En una primera etapa se realizaron análisis con la técnica CHAID a las 11 taxa a partir de los resultados obtenidos y con un marcado interés biológico se decide estudiar 6 de estas

taxa, con la aplicación de otras técnicas, como variables dependientes escogidas una a una:

- **Taxa1**- Archaea, Bacterias, Insectos, Invertebrados, Plantas, Vertebrados no mamíferos, Mamíferos no primate, Primates y Homo Sapiens.
- **Taxa2**- Archaea, Bacterias.
- **Taxa3**- Archaea, Bacterias y Eucariotes.
- Taxa4- Archaea, Bacterias e Invertebrados.
- Taxa5- Insectos y otros invertebrados.
- **Taxa6**- Invertebrados y Vertebrados.
- **Taxa7**- Vertebrados no mamíferos y Mamíferos (mamíferos no primates, primates y homo sapiens).
- Taxa8- Vertebrados no mamíferos y Mamíferos no primates.
- Taxa9- Mamíferos y Primates (homo sapiens).
- Taxa10- Mamíferos no primates, Primates y Homo Sapiens.
- **Taxa11**- Primates y Homo Sapiens.

, 20 variables independientes que representan los aminoácidos:

- 3 clases de 6 tripletes, para los aminoácidos Serina (S), Leucina (L) y Arginina (R).
- 5 clases de 4 tripletes, para los aminoácidos Treonina (T), Alanina (A), Valina (V), Glycina (G) y Prolina (P).
- 2 clases de 3 tripletes, para la Isoleucina (I) y la señal de parada, respectivamente.
- 9 clases de 2 tripletes, para los aminoácidos ácido Glutámico (E), Glutamina (Q), Asparagina (N), ácido Aspártico (D), Histidina (H), Lisina (K), Tirosina (Y), Cisteína (C), y Fenilalanina (F).
- 2 clases de un solo triplete, para la Metionina (M) y el Triptófano (W).

3. LAS DIFERENCIAS EN EL NÚMERO ESTIMADO DE CODONES Y LA CLASIFICACIÓN EVOLUTIVA.

Las especies se clasifican a través de un sistema jerárquico en el cual cada categoría superior incluye otras inferiores. La teoría y la práctica de clasificar los organismos son el objeto de la Taxonomía. Los taxa se pueden clasificar basándose estrictamente en las relaciones de parentesco o valorizando también las novedades adaptativas que aparecen en los linajes. Sin embargo, existe cierta subjetividad en el proceso de clasificación a este nivel. Con el objetivo de eliminar, en alguna medida, la subjetividad presente, la taxonomía no solo se aprovecha de los datos ofrecidos por áreas clásicas de las ciencias biológicas como la Morfología, la Etología, la Citogenética, la Biología Molecular y la Biogeografía, sino además, de las herramientas desarrolladas por la Bioestadística, la Bioinformática y la Informática, las cuales realizan contribuciones significativas a la taxonomía. El análisis taxonómico está estrechamente vinculado con la historia evolutiva de las especies.

Con el propósito de verificar la hipótesis de investigación se aplicaron las técnicas de CHAID y análisis de discriminantes a vectores NEC_k (20-dimensionales) provenientes de las bases de datos descritas en el capítulo 2. El empleo de dos clasificadores diferentes se debe a que la experiencia acumulada en el campo de la bioinformática ha conducido al consenso de que ninguna técnica por separado dará una solución definitiva o muy eficiente a los problemas de clasificación de secuencias de proteínas o de ADN, producto de las indeterminaciones propias de los procesos biológicos y la presencia de muchos ruidos o ausencia de información. La clasificación con el CHAID se ve limitada desde el punto de vista de que cada clasificador que se obtenga, partiendo de algún aminoácido, involucra no a todos los aminoácidos. Sin embargo, a través de este método se pueden detectar cuales aminoácidos y cuales interacciones están asociadas con la clasificación de los vectores NEC_k. Por otra parte, el análisis de discriminante, aunque no incluye el análisis de las interacciones, aporta una verificación alternativa de la hipótesis de investigación y permite evaluar la importancia absoluta de las variables predictivas en la clasificación a través de las correlaciones de estas con la funciones discriminantes, sin importar si la variable se

encuentra o no en las funciones discriminantes. En este capítulo se presentan y discuten los resultados obtenidos utilzando las herramientas mencionadas.

3.1. Comparaciones entre los vectores NEC_k correspondientes a cada taxa

En una primera etapa del análisis se compararon los vectores NEC_k derivados para cada taxa con los correspondientes vectores esperados calculados a partir del código genético señalado para cada grupo taxonómico. Como criterios de comparación se emplearon tres funciones usualmente utilizadas en el análisis comparativo de vectores de probabilidades (o frecuencias):

$$\chi^2: \quad d\left(p_1(x), p_2(x)\right) = \sum_{k=1}^{20} \frac{\left(p_1(x_k) - p_2(x_k)\right)^2}{p_1(x_k)} \tag{3.1.1}$$

Entropía Relativa: $$d\left(p_1(x_k), p_2(x_k)\right) = 2\sum_{k=1}^{20} p_1(x_k) Ln \frac{p_1(x_k)}{p_2(x_k)} \tag{3.1.2}$$

Distancia de Hellinger: $$d\left(p_1(x), p_2(x)\right) = 4\sqrt{\sum_{k=1}^{20}\left(\sqrt{p_1(x_k)} - \sqrt{p_2(x_k)}\right)^2} \tag{3.1.3}$$

Las funciones (3.1.2) y (3.1.3) están expresadas en sus aproximaciones a la función Chi-cuadrado, es decir, la entropía relativa y la distancia de Hellinger han sido multiplicadas por 2, de manera que, si las diferencias entre los vectores que se comparan son suficientemente pequeñas entonces, estas funciones siguen una distribución Chi-cuadrado. En la Tabla 3.1.1 se muestran los resultados de las comparaciones realizadas (utilizando la función 3.1.1) entre los vectores NEC_k y los valores esperados de acuerdo con las estimaciones realizadas a partir de los códigos genéticos correspondientes a cada taxa. Se incluyen, además, los valores de las comparaciones entre los vectores NEC_k calculados a partir de las bases de datos de las secuencias de proteínas y los vectores correspondientes calculados a partir de las bases de datos de las secuencias de genes (derivados de la base de usos de codones). En todas las comparaciones realizadas no se detectaron diferencias estadísticamente significativas entre los vectores. Resultados similares se obtienen para las otras funciones.

Tabla 3.1.1. Resultados de las comparaciones realizadas entre los vectores NEC_k y los valores esperados de acuerdo con las estimaciones realizadas a partir de los códigos genéticos correspondientes a cada taxa.

Grupos de Org. χ^2	**Esperado vs Obs. Uso de codones**	**Esperado vs Obs. Base AA**	**Obs. AA vs Obs. Uso de Codones**
Bacterias	9.43774	10.6452	3.26328
Archaea	17.8872	15.966	3.22017
Plantas	6.40237	5.58239	0.366178
Insectos	6.88278	6.80905	2.35033
invertebrados	10.4236	8.06901	7.8716
vertebrados no mamiferos	6.68555	9.23761	7.30781
Primates	4.42358	4.79673	1.08405
homo sapiens	6.01274	6.03157	2.68314
mamíferos no primates	6.07456	7.44816	4.31209

En este caso se verificó, con todas las funciones utilizadas, que no tenemos criterios estadísticamente suficientes para decir que existen diferencias entre los vectores NECk correspondientes a cada taxa. En otras palabras, en todas las comparaciones realizadas entre vectores, los valores obtenidos (para todas las funciones) son muy pequeños al compararse con el valor de la distribución Chi-cuadrado con 19 grados de libertad (30.1435) y, por lo tanto, siguen una distribución Chi-cuadrado. Los resultados obtenidos pueden observarse en las Tablas 3.1.2, 3.1.3 y 3.1.4. Notemos que las comparaciones entre vectores correspondientes a cada par de taxa analizado dan lugar a valores muy similares de las funciones (3.1.1), (3.1.2) y (3.1.3). Este hecho, pudiera utilizarse en estudios bioinformáticos posteriores, para la elaboración de pruebas de hipótesis o en la implementación de algún nuevo algoritmo.

Tabla 3.1.2. Distribución Chi-cuadrado [19,0.95]= 30.1435

	Archea	**Bact**	**Plantas**	**Invert**	**Insect**	**Vert**	**Mamíf**	**Prim**	**Homo**
Archea	0	0.02045	0.07204	0.08990	0.07097	0.21748	0.17240	0.14214	0.12048
Bact	0.02037	0	0.07217	0.07166	0.06362	0.21259	0.17009	0.15285	0.12858
Plantas	0.06446	0.04787	0	0.01334	0.00582	0.11157	0.06686	0.03586	0.02606
Invert	0.09339	0.05734	0.01303	0	0.00915	0.09024	0.05638	0.04120	0.02965
Insect	0.07129	0.04938	0.00568	0.00913	0	0.08055	0.04729	0.02559	0.02121
Vert	0.27661	0.23412	0.11100	0.09786	0.08667	0	0.00905	0.03328	0.05742
Mamíf	0.20388	0.16946	0.06814	0.06244	0.05028	0.00891	0	0.01385	0.02507
Prim	0.15221	0.12829	0.03516	0.04310	0.02589	0.03335	0.01351	0	0.01334
Homo	0.12255	0.10167	0.02602	0.03207	0.02010	0.05836	0.02615	0.01375	0

Tabla 3.1.3. Comparación de vectores con la Entropía Relativa.

	Archea	**Bact**	**Plantas**	**Invert**	**Insect**	**Vert**	**Mamíf**	**Prim**	**Homo**
Archea	0	0.01931	0.06799	0.08364	0.06633	0.19914	0.18674	0.13197	0.10901
Bact	0.01927	0	0.06193	0.06367	0.05580	0.18612	0.17514	0.12798	0.10782
Plantas	0.06634	0.05448	0	0.01043	0.00439	0.08518	0.07353	0.03107	0.02036
Invert	0.08452	0.05841	0.01039	0	0.00932	0.07902	0.07168	0.03963	0.02683
Insect	0.06682	0.05170	0.00435	0.00926	0	0.06560	0.05692	0.02391	0.01658
Vert	0.21438	0.18980	0.08518	0.08004	0.06684	0	0.00341	0.02446	0.04659
Mamíf	0.19883	0.17674	0.07456	0.07356	0.05824	0.00339	0	0.01905	0.03468
Prim	0.13529	0.12099	0.03080	0.03931	0.02363	0.02447	0.01894	0	0.01378
Homo	0.10931	0.09967	0.02037	0.02739	0.01637	0.04660	0.03500	0.01394	0

Tabla 3.1.4. Construcción de las matrices con la Distancia de Hellinger**.**

	Archea	**Bact**	**Plantas**	**Invert**	**Insect**	**Vert**	**Mamíf**	**Prim**	**Homo**
Archea	0								
Bact	0.0202	0							
Plantas	0.0651	0.0555	0						
Invert	0.0883	0.0612	0.0131	0					
Insect	0.0691	0.0539	0.0057	0.0091	0				
Vert	0.2237	0.2038	0.1062	0.0905	0.0810	0			
Mamíf	0.1756	0.1579	0.0657	0.0579	0.0480	0.0089	0		
Prim	0.1377	0.1293	0.0350	0.0413	0.0254	0.0328	0.0136	0	
Homo	0.1161	0.1080	0.0257	0.0304	0.0204	0.0568	0.0254	0.0135	0

3.2. Construcción de árboles de clasificación mediante el método CHAID atendiendo a las frecuencias de aminoácidos en proteínas

Durante el proceso de evolución molecular que tiene lugar en cada organismo vivo se originan nuevas variantes mutacionales de muchas de las proteínas que conforman el proteoma de este. En el transcurso del tiempo evolutivo la acumulación de mutaciones en genes duplicados deriva en el origen de nuevas especies de organismos, de nuevas proteínas y de nuevas variantes funcionales de proteínas ya existente en las especies ancestros [31].

La aparición de nuevas proteínas en el proceso de especiación pudo conducir a un cambio en la distribución de las frecuencias de aminoácidos. En esta sección proponemos dar respuesta a la primera pregunta de investigación utilizando el método CHAID, el método Discriminante y realizando una evaluación del desempeño de estos clasificadores a través de las curvas ROC y los parámetros calculados a partir de la matriz de confusión.

Para obtener los resultados primeramente se realizó una validación cruzada, una validación al 70% de la muestra inicial, una validación con una muestra externa de 20 vectores de probabilidades, sin tener en cuenta que contienen diferentes tipos de proteínas y solamente guiados por la clasificación inicial de los 9 grupos, también se realizó una validación cruzada y una validación del 70% de la muestra formada por la base llamada extendida, por los resultados obtenidos en el análisis con los 9 grupos se definieron las 6 taxas que reúnen de diferentes formas a los grupos descritos anteriormente.

3.2.1. Aminoácidos asociados con las clasificaciones taxonómicas de organismos vivos.

Como primera etapa en nuestro análisis se aplicó el método CHAID utilizando todas las bases de datos de proteínas descritas en la sección 2.1. En todos los análisis realizados los porcientos de clasificación entre los 9 grupos de organismos no fueron aceptables, con riegos superiores 27.1% en el entrenamiento y 41% en la validación cruzada. Sin embargo, estos análisis nos permitieron detectar que los 20 aminoácidos están asociados con la clasificación taxonómica de las especies y clases analizadas. En la Tabla 3.2.1.1 se muestran los aminoácidos ordenados según sus niveles de significaciones.

El resultado obtenido es esperado desde el punto de vista biológico, si tenemos en cuenta las variaciones en los genomas y proteomas que tuvieron lugar durante el proceso de evolución de los organismos vivos. Como fue explicado en el capítulo 2 la aparición de nuevas especies involucró la aparición de proteínas que no estaban involucradas en procesos esenciales para todos los organismos vivos. De manera que las variaciones en la distribución de aminoácidos deben tender, en general, a ser mayores en la medida que las especies son filogenéticamente más lejanas. Además, si se tienen en cuenta los porcientos que representan algunos genes que codifican para proteínas que están presentes en un número importante de especies, el resultado obtenido es de esperar (ver Fig. 2.1 y 2.2). Luego, la significación estadística de la asociación de los aminoácidos con los taxa debe variar dependiendo de los taxa involucrados en el análisis.

Tabla 3.2.1.1. Significación de los aminoácidos al ser utilizados como variables predictoras en la construcción de árboles de decisión y los porcientos de clasificación alcanzados.

AA	Sig.	%clasificación
Ácido Aspártico	6.7531E-209	69,1
Ácido Glutámico	2.0188E-166	72,4
Triptófano	2.7178E-146	71,6
Valina	1.0947E-132	70,4
Leucina	1.0341E-121	69,8
Arginina	6.9567E-113	69,6
Alanina	1.2499E-108	69,6
Fenilalanina	1.7228E-105	71,6
Metionina	5.9298E-102	71,6
Histeina	5.38196E-92	69,3
Prolina	1.19184E-91	70,4
Aspargina	6.7187E-89	69,8
Isoleucina	1.43677E-84	71,1
Cisteína	1.00026E-80	68,2
Treonina	3.73344E-76	67,8
Lisina	1.0605E-75	68
Sirina	1.34548E-74	66
Glicina	7.06354E-54	68,7
Tirosina	1.1733E-50	68,7
Glutamina	1.83921E-47	69,1

[a]Sig. Significación del estadígrafo de razón verosimilitud Chi-cuadrado. Por simplificación se ha utilizado el simbolismo del SPSS para la notación científica, es decir, por ejemplo, el símbolo E-05 significa 10^{-5}.

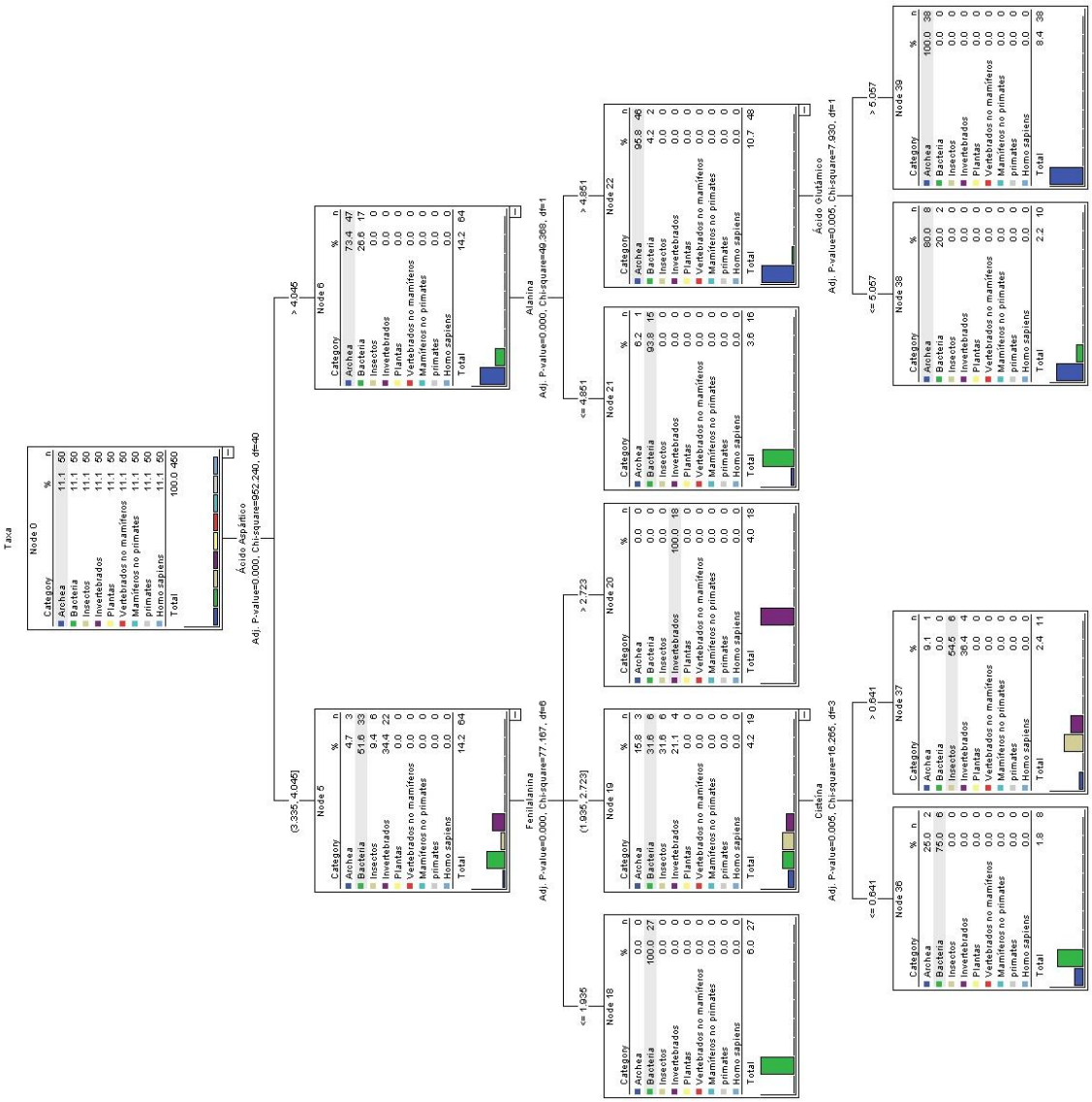


Figura 3.2.1.1A. Sección del árbol de Aminoácidos asociados con los resultados de una validación cruzada en la base curada con las clasificaciones taxonómica de organismos vivos.

Estos resultados sugieren que el análisis se realice en diferentes combinaciones de taxa en las que se reflejen peculiaridades más específicas entre las especies analizadas.

El árbol de clasificación abunda en información biológicamente significativa, ver Anexo 4. En la Fig. 3.2.1.1A se muestra una sección del árbol donde se puede apreciar que si el Ácido Aspártico posee un número esperado de codones (NEC_D) entre 3.335 y 4.045, y si, además, para la Fenilalanina (nodo 5) se cumple que $NEC_F \leq 1.935$ entonces el 100% de las bacterias en este nodo se separa del resto de los taxa. Mientras que si el $NEC_F > 2.723$

entonces el 100% de los invertebrados se separan del resto de los taxa. Por otra parte, si el $NEC_D > 4.045$ y si para la Alanina tenemos $NEC_A \leq 4.851$ entonces el 93.8% de los vectores de distribución corresponde a bacterias. Mientras que si $NEC_A > 4.851$ entonces el 95.8 % de los vectores corresponde a archaeabactarias. Además, si se cumplen las condiciones: $NEC_D > 4.045$, $NEC_A > 4.851$ y $NEC_G > 5.057$ entonces el 100% de los vectores clasificados corresponde a archaeabacterias. Notemos que los NEC_D, NEC_A y NEC_G difieren notablemente de los números esperados de codones en la Tabla 1.2.1 del código genético estándar, hecho que nos sugiere una mayor cercanía a la célula primordial (progenota) ver Anexo 1, para la cual el código genético primitivo pudo encontrarse más alejado del óptimo que los códigos actuales, manifestando valores no optimizados de los *NEC* de estos aminoácidos (ver sección 1.2). Esta observación está en correspondencia con los planteamientos de los autores en [3, 8, 13, 38] expuestos en la sección 1.2. Esta hipótesis biológica explica, además, el porqué existe una separación completa de los eucariotes (resto de los taxa) en estas ramas del árbol, los cuales se encuentran filogenéticamente más distantes del progenota.

En el resto de las ramas del árbol, en las que aparecen los taxa eucariotes, los organismos procariotas están ausentes, lo cual corrobora el hecho de que los NEC_D, NEC_A y NEC_G anteriormente mencionados caracterizan realmente a toda la muestra de procariotas. Sin embargo, estas ramas no aportan una buena clasificación, ver Anexo 4. No obstante, como se muestra en la Tabla 3.2.1.1, el análisis con el método CHAID de todos los taxa revela que los 20 aminoácidos están asociados de forma altamente significativa con la clasificación taxonómica (biológica).

Resultados comparables se obtienen con el análisis de discriminante. En la Tabla 3.2.1.2 se presentan las correlaciones canónicas de las funciones discriminantes canónicas con los taxa y en la Tabla 3.2.1.3 las correlaciones de los aminoácidos con las funciones discriminantes canónicas. En particular, para la mayoría de estas funciones los valores de correlación son altos, indicando un desempeño aceptable de estas funciones en la clasificacicón. Para la base de entrenamiento (70%) se obtuvo un 87% de clasificación correcta, 83% en la validación cruzada y un 80% en la validación externa. De esta manera se verifica, una vez más, la asociación de los aminoácidos con los taxa.

Tabla 3.2.1.2. Correlaciones de las variables discriminantes con las funciones discriminantes canónicas.

Aminoácidos	Función Discriminante							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Serina	-0.666	0.174	0.039	-0.024	-0.180	0.106	0.118	0.234
Alanina	0.284	-0.472	-0.194	-0.313	-0.041	0.030	-0.083	0.384
Leucina	0.034	0.029	-0.504	-0.120	-0.026	0.013	0.348	-0.369
Asparagina	-0.162	-0.020	0.461	0.329	-0.069	0.212	-0.371	-0.206
Triptófano	0.004	-0.142	-0.416	0.084	-0.337	-0.191	-0.079	-0.144
Ácido Aspártico	0.060	-0.227	0.396	-0.174	0.240	0.167	0.095	-0.176
Tirosina	0.051	0.159	0.137	0.410	-0.204	0.240	0.008	-0.247
Isoleucina	0.231	0.390	0.178	0.103	-0.394	0.030	0.107	-0.361
Prolina	-0.107	-0.088	-0.115	-0.126	0.373	0.166	-0.153	0.292
Ácido Glutámico	0.180	0.293	0.170	-0.162	0.345	-0.166	-0.163	0.155
Treonina	-0.096	-0.285	-0.114	0.312	0.144	-0.407	0.106	0.143
Valina	0.361	0.011	-0.049	-0.094	-0.344	0.150	0.505	-0.058
Metionina	0.040	0.096	0.247	0.065	-0.047	-0.318	0.488	-0.063
Lisina	0.060	0.273	0.281	-0.037	0.010	0.097	-0.292	-0.253
Glicina	0.089	-0.132	0.011	-0.138	0.063	-0.187	-0.240	-0.145
Glutamina	-0.317	-0.122	-0.208	0.399	0.334	-0.004	-0.068	0.459
Arginina	0.064	-0.121	-0.104	-0.159	0.065	-0.279	-0.032	0.435
Fenilalanina	0.029	0.156	-0.005	0.146	-0.303	0.144	0.117	-0.404
Cisteína	-0.191	0.108	-0.027	0.170	-0.049	0.205	0.143	-0.276
Histidina	-0.147	-0.143	-0.022	0.162	0.038	-0.103	0.027	0.273

Sin embargo, estos hechos nos sugieren realizar un análisis agrupando los taxa siguiendo criterios biológicos con el propósito de alcanzar una mayor significación estadística en la diferenciación de los taxa. En particular, por su importancia biológica, se consideran los grupos taxonómicos que divergen de un ancestro común.

Tabla 3.2.1.3. Funciones discriminantes canónicas obtenidas con la introducción de todos los aminoácidos que satisfacen el test de tolerancia y con el método Stepwise.

Amino	1		2		3		4		5		6		7		8	
ácidos	Todas	Stpws	Todas	Stpws	Todas	Stpws	Todas	Stpws	Todas	Stpws	Todas	Stpws	Todas	Stpws	Todas	Stpws
Ala	0.632	0.630	1.990	-0.905	0.146	0.177	1.449	-0.037	0.984	-0.304	0.398	1.037	0.434	0.385	0.762	1.556
Cys	1.512	-0.233	0.592	0.488	0.130	0.161	0.327	1.088	0.425	0.239	0.346	1.140	-0.162	0.980	-0.425	0.368
Asp	1.637	-0.332	2.772	-1.711	1.225	1.257	2.051	-0.708	-0.061	0.753	0.088	1.664	-0.917	1.640	-1.126	-0.313
Glu	0.351	0.880	-0.467	1.574	0.035	0.066	1.057	0.436	-0.270	0.954	0.872	0.561	0.660	0.230	1.033	1.810
Phe	1.267	0.035	1.384	-0.321	-0.449	-0.417	0.235	1.137	0.597	0.051	-0.022	1.573	-0.004	0.771	0.518	1.324
Gly	1.244		1.100		-0.031		1.444		0.763		1.367		0.880		-0.780	
His	0.367	0.999	1.050	-0.029	-0.469	-0.437	0.261	0.978	1.249	-0.410	1.159	0.496	-0.001	0.709	0.709	1.533
Ile	-0.217	1.463	0.261	0.834	0.058	0.090	1.125	0.327	1.263	-0.589	0.560	0.749	0.463	0.406	0.313	1.094
Lys	1.441		0.979		-0.032		1.101		0.999		1.934		0.597		-0.854	
Leu	0.997	0.285	1.284	-0.209	-0.824	-0.792	1.577	-0.192	-0.184	0.948	1.055	0.575	-0.048	0.847	-0.979	-0.182
Met	1.497	-0.195	1.556	-0.493	3.228	3.260	1.341	-0.027	0.054	0.985	3.357	-1.724	-2.075	2.932	-0.047	0.735
Asn	0.609	0.630	2.218	-1.120	1.509	1.540	0.397	1.068	0.621	0.011	0.316	1.040	1.550	-0.693	0.038	0.825
Pro	-0.015	1.340	0.316	0.732	-0.189	-0.158	1.268	0.068	-0.669	1.365	0.169	1.648	0.456	0.249	0.403	1.225
Gln	1.607	-0.328	1.409	-0.332	-0.659	-0.628	-0.450	1.869	0.333	0.296	0.106	1.369	0.221	0.594	0.339	1.133
Arg	0.862	0.465	0.966	0.082	0.335	0.367	1.154	0.146	1.069	-0.229	1.396	0.186	0.214	0.549	0.136	0.946
Ser	2.636	-1.368	0.624	0.466	0.257	0.288	2.142	-0.733	1.393	-0.658	0.837	0.554	0.167	0.669	0.469	1.258
Thr	0.576	0.771	1.777	-0.747	-0.112	-0.080	-0.388	1.638	-0.060	1.096	3.200	-1.489	-0.129	0.920	-0.220	0.584
Val	0.022	1.327	0.704	0.327	0.038	0.070	0.451	0.826	1.382	-0.647	0.318	1.292	-0.468	1.179	0.454	1.274
Trp	1.538	-0.154	0.928	0.086	-1.325	-1.293	0.982	0.208	3.371	-2.464	1.523	-0.087	2.649	-1.953	-0.319	0.516
Tyr		1.264		1.085		0.031		1.451		0.606		1.483		0.830		0.789
(Const)	-54.96	-24.16	-65.82	0.80	-5.44	-7.37	-61.69	-21.39	-34.66	-11.50	-56.01	-38.84	-12.55	-35.65	-0.64	-49.52

3.2.2. Aminoácidos asociados con la clasificación taxonómica en archaeabacterias, bacterias y eucariotes.

Los resultados obtenidos en la clasificación de organismos vivos nos sugiere limitar nuestro campo de análisis para la clasificación de los tres reinos: archaea, bacterias y eucariotes, partiendo de una base curada y realizando una validación cruzada como se muestra en la Tabla 3.2.2.1, se mantiene para este taxa la asociación entre los 20 aminoácidos. Los porcientos de clasificación aumentan considerablemente lo cual sugiere la ya demostrada hipótesis de la existencia de los tres reinos bien definidos que forman el árbol filogenético universal ver Anexo1. En la Tabla 3.2.2.1 se observa que el Ácido Aspártico posee la mejor significación para la ramificación del nodo inicial entre todos los aminoácidos y los resultados del árbol de clasificación son aceptables (ver Tabla 3.2.2.2 y Figura 3.2.2.1). Sin embargo, la Cisteína, aunque posee una menor significación, alcanza 99,3 % de clasificación.

Tabla 3.2.2.1. Significación de los aminoácidos al ser utilizados como variables predictoras en la construcción de árboles de decisión y los porcientos de clasificación alcanzados.

AA	Sig.	%clasificación
Alanina	1.13359E-80	98
Cisteína	5.12355E-70	99,3
Ácido Aspártico	1.926E-137	98
Ácido Glutámico	1.6169E-121	98
Fenilalanina	3.35584E-86	98,4
Glicina	9.26829E-42	98,4
Histeina	2.04308E-93	97,6
Isoleucina	3.51131E-27	97,1
Licina	2.1581E-34	96,9
Leucina	2.39256E-60	97,6
Metionina	1.12471E-62	97,8
Aspargina	2.42534E-40	99,3
Prolina	1.02933E-76	96,7
Glutamina	3.86159E-21	97,3
Arginina	1.24853E-63	97,8
Serina	4.71165E-78	98,4
Treonina	1.68436E-17	96,9
Valina	6.4982E-98	98,7
Triptófano	1.1792E-101	98
Tirosina	4.6616E-46	97.3

[a]Sig. Significación del estadígrafo de razón verosimilitud Chi-cuadrado. Por simplificación se ha utilizado el simbolismo del SPSS para la notación científica, es decir, por ejemplo, el símbolo E-05 significa 10^{-5}.

Tabla 3.2.2.2. Clasificación obtenida con método CHAID en la bases de datos curada con validación cruzada.

Muestras Observadas	Predicciones			
	Archaeas	Bacterias	Eucariotes	Exactitud
Archaeas	46	3	1	92.0%
Bacterias	2	48	0	96.0%
Eucariotes	0	1	349	99.7%
% Total	10.7%	11.6%	77.8%	98.4%

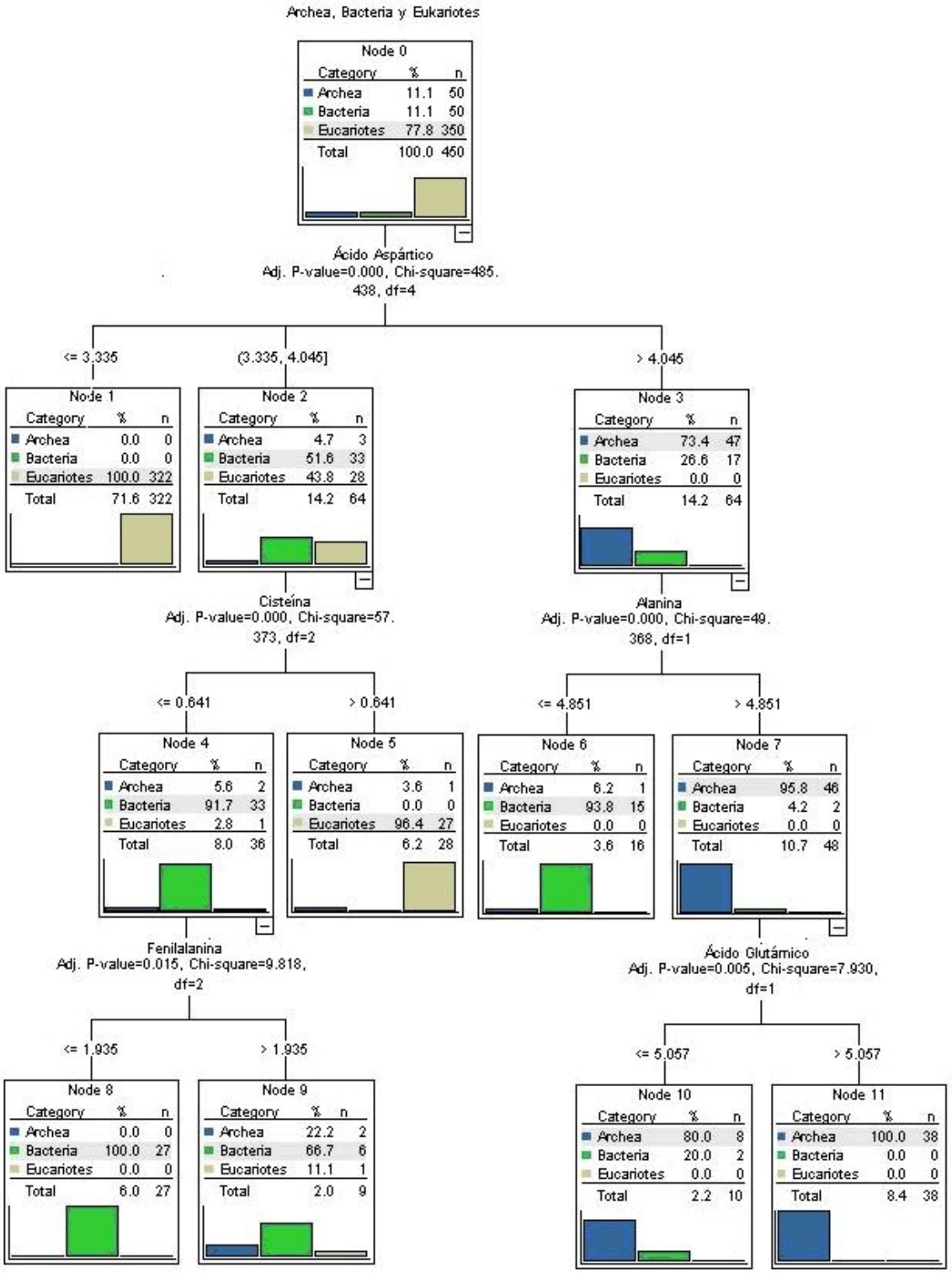


Figura 3.2.2.1. Árbol de Aminoácidos asociados con los resultados en la base de datos curada con validación cruzada en las clasificaciones taxonómicas de archaea, bacterias y eucariotes.

Tabla 3.2.2.3. Clasificación obtenida con método CHAID en la bases de datos extendida con validación cruzada.

Muestra Observada	Predicciones			
	Archaeas	Bacterias	Eucariotes	Exactitud
Archaeas	65	5	0	92.9%
Bacterias	1	69	0	98.6%
Eucariotes	1	3	487	99.2%
% Total	10.6%	12.2%	77.2%	98.4%

Además de esto, el método CHAID nos permite detectar los aminoácidos que interaccionan en esta clasificación, lo cual resulta de gran interés desde el punto de vista biológico, pues destaca el papel de conjunto jugado por los aminoácidos en la diferenciación de los taxa. En otras palabras, se han detectado interaciones estadísticamente significativas entre los aminoácidos, las cuales son, además, biológicamente significativas, pues permiten derivar reglas de clasificación capaces de diferenciar los taxa (ver Anexo 5). Se puede realizar una discusión más abundante acerca de este interesante tema pero está fuera del alcance y del objetivo de este trabajo.

Realizando una validación cruzada a la base de datos extendida los procientos de clasificación son igualmente buenos, lo cual mostramos en la Tabla 3.2.2.3, mientras en el árbol (ver Anexo 5) podemos observar que para esta base de datos el aminoácido con mayor significación es la Histidina y así aparece en el nodo principal.

La base usada en el análisis anterior se incrementó en el número de secuencias y se formaron 200 nuevos vectores NEC_K para cada taxa. En esta base se seleccionó aleatoriamente el 70% de los datos como entrenamiento y el resto para validación externa con el propósito de aplicar, además de la técnica CHAID, el análisis de discriminante y comparar el desempeño de estos. En esta ocasión el aminoácido Serina fue el de mayor significación para la ramificación del nodo inicial del árbol. En la Tabla 3.2.2.4 se muestran los resultados para este árbol.

Tabla 3.2.2.4. Clasificación obtenida con método CHAID en la nueva base de datos extendida tomando aleatoriamente el 70% de la base como entrenamiento y el resto usado en validación externa. El árbol inicia su ramifiación con el aminoácido Serina.

		Predicciones			
muestra	observadas	archaea	bacteria	eucariotes	Exactitud
entrenamiento	archaeas	138	10	0	93.2%
	bacterias	6	121	7	90.3%
	eucariotes	2	4	126	95.5%
	% Total	35.3%	32.6%	32.1%	93.0%
validac	archaeas	49	3	0	94.2%
	bacterias	6	55	5	83.3%
	eucariotes	2	5	61	89.7%
	% Total	30.6%	33.9%	35.5%	88.7%

3.2.2.1. Análisis de Discriminante y la evaluación del desempeño de los clasificadores.

Con el análisis de discriminante realizado en esta taxa se corroboró el resultado, previamente obtenido con el CHAID, de que todos los aminoácidos están asociados con la clasificación de los vectores NEC_k en los tres reinos. En la Tabla 3.2.2.1.1 se puede ver que, incluso, aminoácidos como la Isoleucina y la Asparagina, los cuales no están incluidos en las combinaciones lineales de las funciones discriminantes cuando se utiliza el método Stpwise, poseen correlaciones mayores que algunos de los incluídos. La explicación de este hecho se encuentra en la matríz de correlaciones mostrada en el Anexo 7, se puede apreciar que la Isoleucina (r_1 = -0.159 y r_2 = 0.38, en la Tabla 3.2.2.1.1), posee coeficientes de correlación absolutos elevados y altamente significativos ($p < 0.01$) con la Glutamina (-0.768), la Arginina (-0.627), Alanina (-0.567) y la Lisina (-0.506), las cuales se incluyen en las funciones discriminantes (Tabla 3.2.2.1.2). Luego, si no se aplica un método Stepwise para la introducción de las variables entonces debemos esperar que todas las variables, que superen el test de tolerancia, esten presentes en la combinación lineal que conforman las funciones discriminantes, a pesar de que estás últimas incluirán información redundante, la cual se evidencia en las correlaciones existentes entre las mismas. En la Tabla 3.2.2.1.2 se presentan las funciones discriminantes obtenidas por el método Stepwise minimizando la Lambda de Wilk y sin aplicar este método.

Tabla 3.2.2.1.1. Correlaciones de las variables discriminantes con las funciones discriminantes canónicas.

Aminoácido	Función discriminante	
	1	2
Histidina	0.352*	-0.011
Ácido Aspártico [a]	-0.278*	-0.209
Ácido Glutámico	-0.274*	-0.009
Triptófano [a]	0.243*	-0.023
Valina	-0.200*	-0.193
Prolina [a]	0.200*	-0.166
Leucina [a]	0.180*	0.038
Treonina [a]	0.175*	0.034
Metionina	0.158*	0.080
Alanina	-0.058	-0.497*
Arginina	-0.012	-0.471*
Glutamina	0.191	-0.429*
Tirosina	-0.055	0.391*
Isoleucina [a]	-0.159	0.380*
Serina	0.350	0.377*
Asparagina [a]	-0.159	0.343*
Lisina	-0.262	0.338*
Cisteína	0.240	0.263*
Fenilalanina [a]	0.160	0.186*
Glicina	0.000	-0.042*

* La mayor correlación absoluta entre cada variable y las funciones discriminantes obtenidas por el método Stepwise minimizando la Lambda de Wilk.
[a] Aminoácidos que no se incluyen en la combinación lineal de variables de las funciones discriminantes.

Mientras, en la Tabla 3.2.2.1.3 se puede apreciar que la eficacia de las funciones discriminates en la separación de los casos en grupos, expresada a través de las correlaciones canónicas, es similar para ambos procedimientos. Además, los valores de la Lambda de Wilk y la significación del test Chi-cuadrado indican que las capacidades discriminatorias de las funciones obtenidas por estos procedimientos son similares. En particular, para todas las funciones los valores de estos parámetros son altos, indicando el buen desempeño de las funciones discriminantes.

Los resultados de la clasificación global no son estadísticamente diferentes para los métodos de obtención de las funciones discriminantes y para el método CHAID.

Tabla 3.2.2.1.2. Funciones discriminantes canónicas obtenidas con la introducción de todos los aminoácidos que satisfacen el test de tolerancia y con el método Stepwise.

Aminoácido	Todas		Stepwise	
	Función discriminante		Función discriminante	
	1	2	1	2
Alanina	0.817868518	0.557817117	-0.634610589	-0.073293422
Cisteína	3.310732231	-1.692376434	1.901286543	1.965185789
Ácido Aspártico	1.558358609	0.317792496	-	-
Ácido Glutámico	0.792671149	-0.964414539	-0.747640779	1.387547125
Fenilalanina	1.464973618	0.261072156	-	-
Glicina	0.905491907	0.298178086	-0.661236016	0.17008811
Histidina	3.650391437	-0.681412275	2.111016944	1.442137596
Isoleucina	1.251241606	0.315105464	-	-
Lisina	1.291028255	0.781428574	-0.259736438	-0.365299324
Leucina	1.692656627	0.627164431	-	-
Metionina	2.232989862	0.355582763	0.735120813	0.096567728
Asparagina	1.755662058	0.466027904	-	-
Prolina	1.680706246	0.694573323	-	-
Glutamina	2.661642231	1.967428169	1.217848927	-1.606940573
Arginina	1.453948483	0.854261427	0.061374096	-0.534533261
Serina	2.386767639	-0.472410556	0.987602167	0.817741217
Treonina	1.355212891	0.56985644	0.924887649	-0.466115737
Valina	2.522376259	1.000621086	-	-
Tirosina	-	-	-1.546234127	0.906075053
Triptófano	1.599846115	-0.626305971	-	-
(Constante)	-95.07061812	-21.02880013	-2.504492874	-6.918114573

Tabla 3.2.2.1.3. Eficacia de las funciones discriminantes a través de las correlaciones canónicas y los valores de la Lambda de Wilk.

Función	Valor principal	% de Varianza	% Var. Acum.	Corr. Canónica	Función	Lambda de Wilks	Chi cuadrado	g.l.	Sig.
Stepwise									
1	7.052	78.88	78.88	0.936	1 a 2	0.043	1275.937	24	0.000
2	1.888	21.12	100	0.809	2	0.346	430.085	11	0.000
Todas las variables									
1	7.248	78.578	78.578	0.937	1 a 2	0.041	1286.631	38	0.000
2	1.976	21.422	100	0.815	2	0.336	438.414	18	0.000

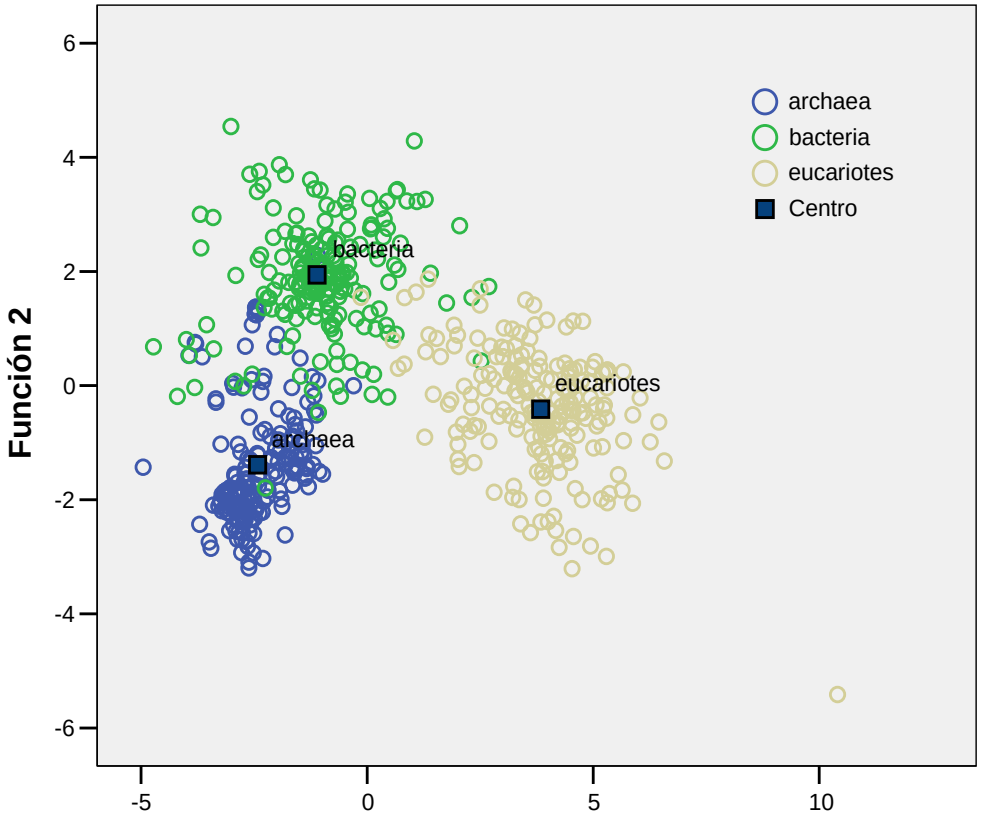


Figura 3.2.2.1.1 Gráfico de dispersión de la función Discriminante.

Este hecho se ilustra en las curvas ROC obtenidas (Figura 3.2.2.1.2) y en la Tabla 3.2.2.1.4, en la que se muestra que los intervalos de confianza asintóticos para 95% de confianza de las áreas bajo la curva ROC se solapan. Sin embargo, al utilizar los parámetros derivados de la matríz de confusión para evaluar el desempeño de estos clasificadores, nos sugieren que existen algunas diferencias entre los clasificadores. En la Tabla 3.2.2.1.5 se muestran los valores de los parámetros mencionados.

Tabla 3.2.2.1.4. Resultado del área bajo la curva en los tres métodos utilizados.

Resultados del Análisis	Área	Error Estándar	Sig. Asintótica	Intervalo de confianza asintótico para el 95%	
				Límite superior	Límite inferior
Probabilidad Archaea (Análisis Disc. Stepwise)	0.991	0.003	0.000	0.985	0.996
Probabilidad Archaea (Análisis CHAID)	0.987	0.004	0.000	0.980	0.995
Probabilidad Archaea (Análisis Disc. Todas)	0.991	0.003	0.000	0.986	0.997
Probabilidad Bacteria (Análisis Disc. Stepwise)	0.982	0.004	0.000	0.974	0.990
Probabilidad Bacteria (Análisis CHAID)	0.967	0.008	0.000	0.952	0.983
Probabilidad Bacteria (Análisis Discriminante)	0.983	0.004	0.000	0.975	0.991
Probabilidad Eucariotes (Análisis Disc. Stepwise)	0.999	0.001	0.000	0.998	1.000
Probabilidad Eucariotes (Análisis CHAID)	0.984	0.005	0.000	0.975	0.994
Probabilidad Eucariotes (Análisis Discriminante)	0.999	0.001	0.000	0.997	1.000

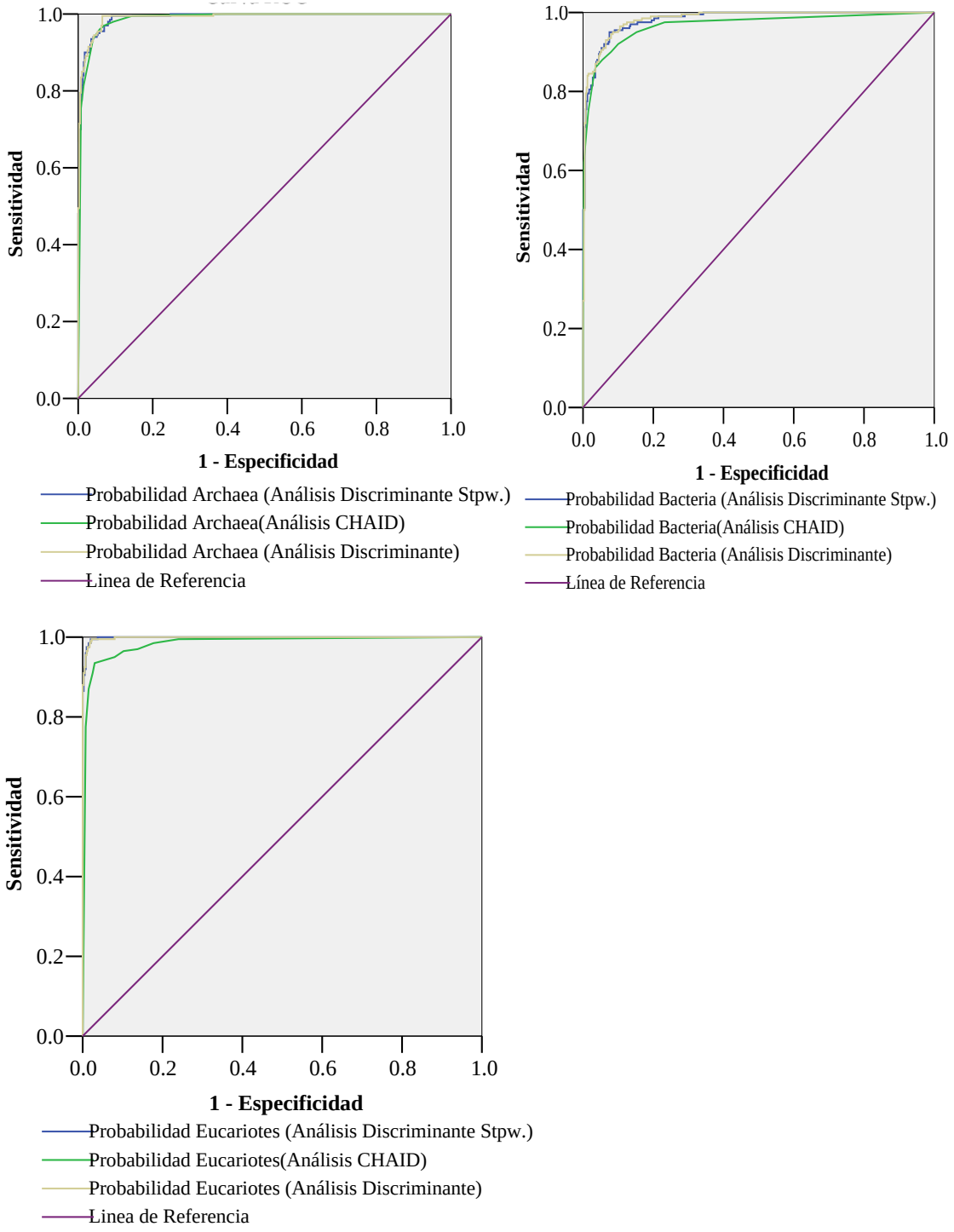


Figura 3.2.2.1.2 Curvas ROC obtenidas con los dos métodos de discriminante y con el método CHAID.

Tabla 3.2.2.1.5 Parámetros calculados a partir de la matriz de confusión para evaluar el desempeño de los clasificadores utilizados.

Org.	Archaea	Bacteria	Eucariotes	Razón de TP	Razón TN	Precisión
Predicciones de los miembros del Grupo con Anl. Discriminante Stepwise						
70 % base de datos extendida						
Archaea	93.2	6.8	0.0	93.2	94.0	92.6
Bacteria	8.2	90.3	1.5	90.3	95.4	90.3
Eucariotes	0.0	2.3	97.7	97.7	91.8	98.5
Validación cruzada						
Archaea	92.6	7.4	0.0	92.6	92.5	91.3
Bacteria	9.7	88.8	1.5	88.8	94.3	88.1
Eucariotes	0.0	3.8	96.2	96.2	90.8	98.4
Validación externa						
Archaea	88.5	11.5	0.0	88.5	92.5	93.9
Bacteria	4.5	92.4	3.0	92.4	90.8	84.7
Eucariotes	0.0	7.4	92.6	92.6	90.7	96.9
Predicciones de los miembros del Grupo con Anl. Discriminante (todas)						
70 % base de datos extendida						
Archaea	91.9	8.1	0.0	91.9	94.4	93.2
Bacteria	7.5	91.0	1.5	91.0	94.6	89.1
Eucariotes	0.0	2.3	97.7	97.7	91.5	98.5
Validación cruzada						
Archaea	91.9	8.1	0.0	91.9	92.5	92.5
Bacteria	8.2	90.3	1.5	90.3	93.2	86.4
Eucariotes	0.0	5.3	94.7	94.7	91.1	98.4
Validación externa						
Archaea	90.4	9.6	0.0	90.4	91.8	92.2
Bacteria	6.1	90.9	3.0	90.9	91.7	85.7
Eucariotes	0.0	7.4	92.6	92.6	90.7	96.9
Predicciones de los miembros del Grupo con CHAID						
70 % base de datos extendida						
Archaea	95.3	4.7	0.0	95.3	95.5	99.3
Bacteria	0.7	95.5	3.7	95.5	95.4	90.8
Eucariotes	0.0	4.5	95.5	95.5	95.4	96.2
Validación externa						
Archaea	47.0	4.0	1.0	90.4	91.8	100.0
Bacteria	0.0	62.0	4.0	93.9	90.0	84.9
Eucariotes	0.0	7.0	61.0	89.7	92.4	92.4

3.2.3. Aminoácidos asociados con la clasificación taxonómica en archaeabacterias y bacterias.

En un primer análisis se utilizó la técnica del CHAID con validación cruzada en la base de datos curada. El método CHAID construye, por defecto, el árbol de la variable con mayor significación estadística. Para esta base el aminoácido de mayor significación es la Alanina, cuyo árbol se muestra en la Figura 3.2.3.1. En la tabla de clasificación correspondiente se aprecia que para las bacterias se alcanza un 100% de clasificación, mientras que en la clasificación total se logra un 96% (Tabla 3.2.3.1). No obstante, en la Tabla 3.2.3.2 se puede ver que todos los aminoácidos están fuertemente asociados con la clasificación taxonómica biológica y que el aminoácido con mayor significación estadística no es el que causa el mejor porciento de clasificación.

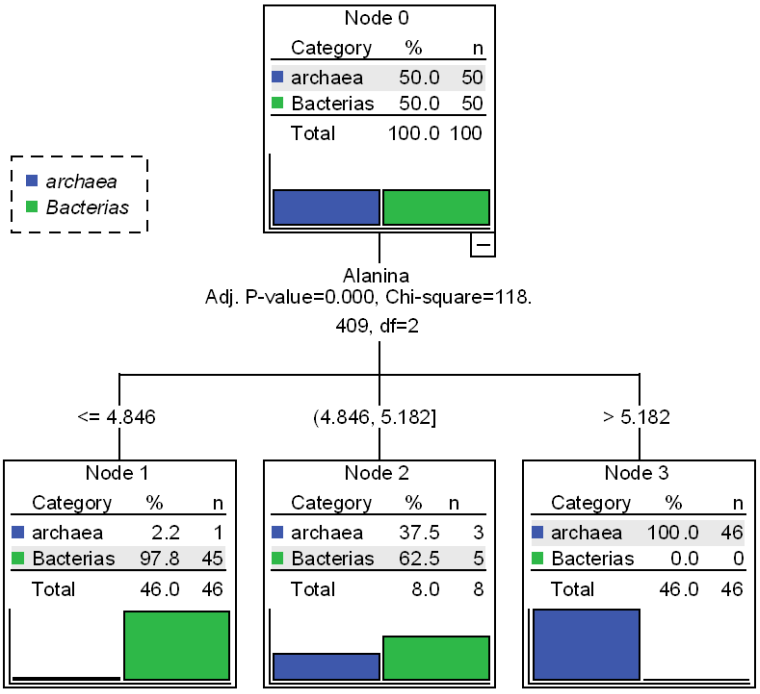


Figura 3.2.3.1 Árbol de Aminoácidos asociados con los resultados en la base de datos curada con validación cruzada en las clasificaciones taxonómicas de archaea y bacterias.

Tabla 3.2.3.1. Clasificación obtenida con método CHAID en la bases de datos curada con validación cruzada.

Muestras Observadas	Prediciones		
	archaea	Bacterias	Exactitud
archaea	46	4	92.0%
Bacterias	0	50	100.0%
% Total	46.0%	54.0%	96.0%

Tabla 3.2.3.2. Significación de los aminoácidos al ser utilizados como variables predictoras en la construcción de árboles de decisión y los porcientos de clasificación alcanzados.

AA	Sig.	%clasificación
Serina	0.043322937	99
Glicina	0.042580131	96
Fenilalanina	0.005793191	100
Valina	0.002078548	97
Cisteína	0.001020605	96
Prolina	0.000393157	98
Treonina	0.000319579	97
Glutamina	0.000292468	97
Tirosina	0.000151604	93
Metionina	2.52667E-08	94
Leucina	9.64362E-10	93
Arginina	4.37685E-10	92
Triptófano	2.64469E-13	96
Ácido Glutámico	1.63794E-14	93
Histidina	1.4771E-14	97
Isoleucina	5.9E-15	95
Ácido Aspártico	7.09697E-19	96
Lisina	6.90823E-20	96
Asparagina	6.90823E-20	97
Alanina	1.28E-24	96

[a]Sig. Significación del estadígrafo de razón verosimilitud Chi-cuadrado. Por simplificación se ha utilizado el simbolismo del SPSS para la notación científica, es decir, por ejemplo, el símbolo E-05 significa 10^{-5}

En la Figura.3.2.3.2 se muestra el árbol de decisión obtenido por método CHAID en la base de datos curada forzando la entrada del aminoácido Fenilalanina, mientras que en la Tabla 3.2.3.3 se muestra los porcientos de clasificación. Los resultados muestran que forzando la entrada del aminoácido Fenilalanina produce una separación definitiva de las archaeas y bacterias. Notemos que, para este aminoácido se obtiene el 100 % de

clasificación, aunque está lejos de tener una buena significación estadística (si lo comparamos con el resto de los aminoácidos).

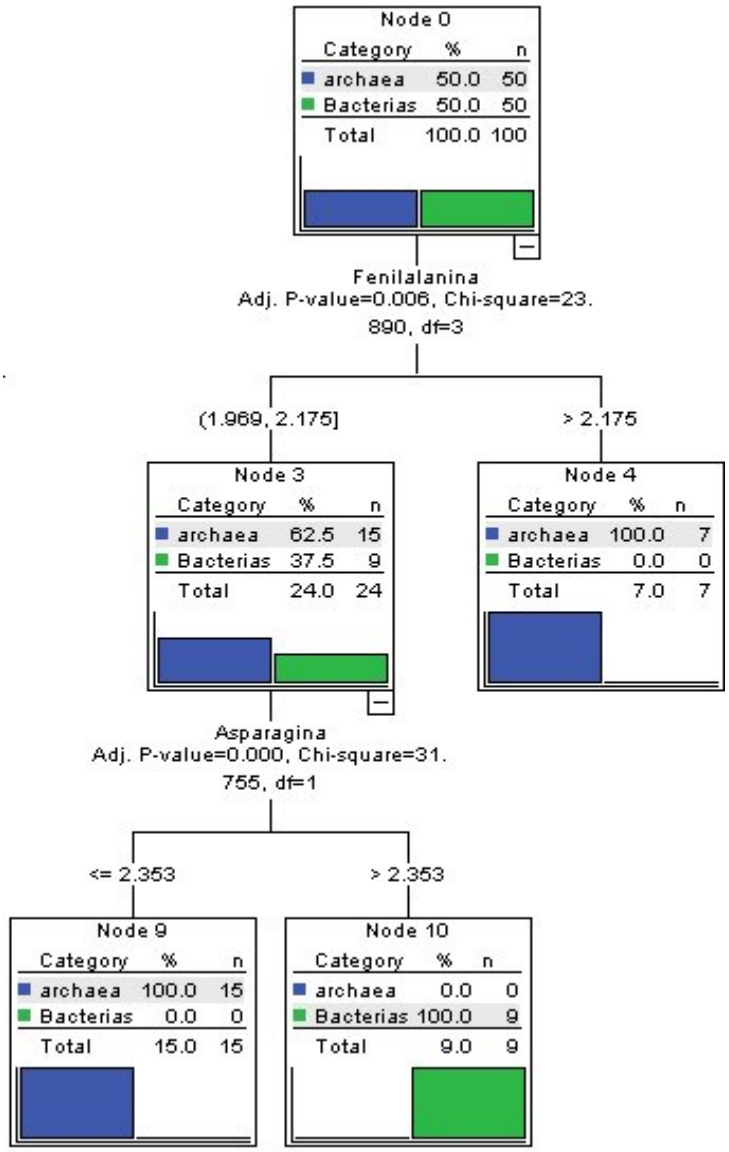


Figura 3.2.3.2.A. Árbol de Aminoácidos asociados con los resultados en la base de datos curada con validación cruzada en las clasificaciones taxonómicas de archaea y bacterias, forzando la Fenilalanina.

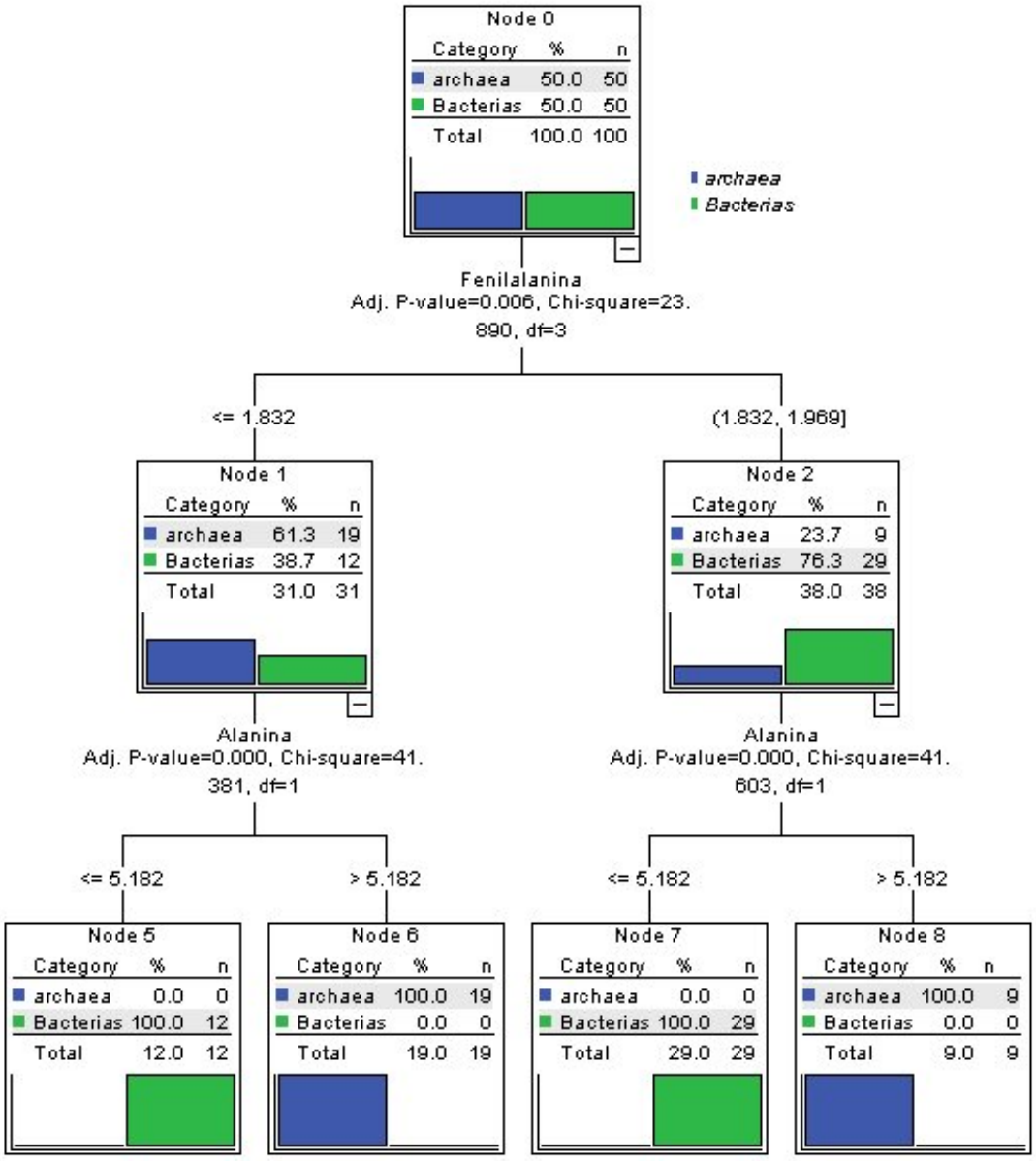


Figura 3.2.3.2.B. Árbol de Aminoácidos asociados con los resultados en la base de datos curada con validación cruzada en las clasificaciones taxonómicas de archaea y bacterias, forzando la Fenilalanina.

Tabla 3.2.3.3. Clasificación obtenida con método CHAID en la bases de datos curada con validación cruzada, forzando la Fenilalanina.

Muestras Observadas	Predicciones		
	archaea	Bacterias	Exactitud
archaea	50	0	100.0%
Bacterias	0	50	100.0%
% total	50.0%	50.0%	100.0%

Tabla 3.2.3.4. Clasificación obtenida con método CHAID en la nueva base de datos extendida tomando aleatoriamente el 70% de la base como entrenamiento y el resto usado en validación externa.

		Predicciones		
muestra	Observada	archaea	bacteria	Exactitud
entrenamiento	archaea	148	1	99.3%
	bacteria	5	141	96.6%
	%Total	51.9%	48.1%	98.0%
prueba	archaea	51	0	100.0%
	bacteria	1	53	98.1%
	%Total	49.5%	50.5%	99.0%

Hasta este punto, se ha verificado que es posible diferenciar los reinos de bacterias y archaeas utilizando bases de secuencias en las que se ha reducido el número de secuencias que comparten características comunes a ambas taxa y expresan, en mayor medidad, la variabilidad propia de cada taxa. Sin embargo, por construcción, la base curada no contiene la variabilidad necesaria, en las secuencias de proteínas que la conforman, para ser útil como base de entrenamiento que permita obtener un clasificador capáz de alcanzar un buen desempeño ante una base externa con alta variabilidad de secuencias. Estos hechos evidencian que si se desea clasificar secuencias de proteínas con mayor variabilidad en las distribuciones de aminoácidos correspondientes, es necesario extender la base curada con secuencias que compartan características estadísticas comunes a ambos taxa. En la Tabla 3.2.3.4 se muestra el resultado del análisis con el CHAID de la base extendida formada por 200 vectores de cada taxa. En la base de entrenamiento (70% de la base) se alcanzó el 98% de clasificación total, mientras que en la validación externa el 99% (30% de la base).

3.2.3.1. Análisis de Discriminante y la evaluación del desempeño de los clasificadores.

El análisis de discriminante realizado en esta taxa muestra que todos los aminoácidos están asociados con la clasificación de los vectores NEC_k. En la Tabla 3.2.3.1.1 se puede ver que, incluso aquellos que no están presentes en uno de los dos métodos Tabla 3.2.3.1.2, o en ambos, como es el caso de la Tirosina poseen correlaciones mayores que algunos de los incluidos.

En la Tabla 3.2.3.1.3 se puede apreciar que los valores de la Lambda de Wilk y la significación del test Chi-cuadrado indican que las capacidades discriminatorias de las funciones obtenidas por estos procedimientos son similares. La eficacia de las funciones discriminantes en la separación de los casos en grupos, se expresa a través de los valores de las correlaciones canónicas.

Para evaluar el desempeño del CHAID y el Discriminante usamos las curvas ROC obtenidas (Figura 3.2.3.1.1) y en la Tabla 3.2.3.1.4, tenemos los valores de las áreas bajo la curva, estos elementos muestran que no hay diferencias significativas entre los dos métodos. Al utilizar los parámetros derivados de la matríz de confusión, nos sugieren que las diferencias entre los clasificadores son mínimas. En la Tabla 3.2.3.1.5 se muestran los valores de los parámetros mencionados.

Tabla 3.2.3.1.1. Correlaciones de las variables discriminantes con las funciones discriminantes canónicas.

Aminoácido	**Función Discriminante**
Glutamina	0.55
Isoleucina [a]	-0.45
Arginina	0.428
Lisina	-0.403
Tirosina [a]	-0.367
Alanina	0.357
Prolina	0.318
Asparagina	-0.311
Serina	-0.246
Fenilalanina [a]	-0.202
Histidina [a]	0.138
Cisteína [a]	-0.118
Leucina	0.104
Ácido Aspártico	0.087
Valina	0.067
Treonina	0.045
Triptófano [a]	0.039
Ácido Glutámico [a]	-0.029
Metionina	-0.025
Glicina	0.009

Tabla 3.2.3.1.2. Funciones discriminantes canónicas obtenidas con la introducción de todos los aminoácidos que satisfacen el test de tolerancia y con el método Stepwise.

Aminoácidos	Función discriminante	
	Todas	Stepwise
Alanina	0.86593423	0.8051423
Cisteína	0.56531521	-
Ácido Aspártico	0.82557749	0.71594929
Ácido Glutámico	0.33801189	-
Fenilalanina	0.10955719	-
Glicina	0.55514944	0.64425764
Histidina	0.26960293	-
Isoleucina	0.26424837	-
Lisina	1.55534119	1.54844199
Leucina	1.85565316	1.99078812
Metionina	2.25971271	2.38887696
Asparagina	2.20713337	2.33146195
Prolina	2.27098881	2.46609259
Glutamina	2.69197518	2.63442816
Arginina	1.15922745	1.12068112
Serina	0.47466926	0.50092619
Treonina	2.20243491	2.44662457
Valina	1.10444373	1.04160686
Triptófano	0.05608013	-
(Constant)	59.0681373	61.4489587

Tabla 3.2.3.1.3. Eficacia de las funciones discriminantes a través de las correlaciones canónicas y los valores de la Lambda de Wilk.

Función	Valor principal	% de Varianza	% Var. Acum.	Corr. Canónica	Función	Lambda de Wilks	Chi cuadrado	g.l.	Sig.
Stepwise									
1	3.789	100	100	0.889	1	0.209	448.777	13	0.000
Todas las variables									
1	3.886	100	100	0.892	1	0.205	449.715	19	0.000

Tabla 3.2.3.1.4. Resultado del área bajo la curva en los tres métodos utilizados.

Resultados del Análisis	Área	Error Estándar	Sig. Asintótica	Intervalo de confianza asintótico para el 95%	
				Límite superior	Límite inferior
Probabilidad Bacteria (Análisis Disc. Stepwise)	0.995	0.002	0.000	0.991	0.999
Probabilidad Bacteria (Análisis CHAID)	0.996	0.002	0.000	0.992	1.000
Probabilidad Bacteria (Análisis Discriminante)	0.996	0.002	0.000	0.992	1.000

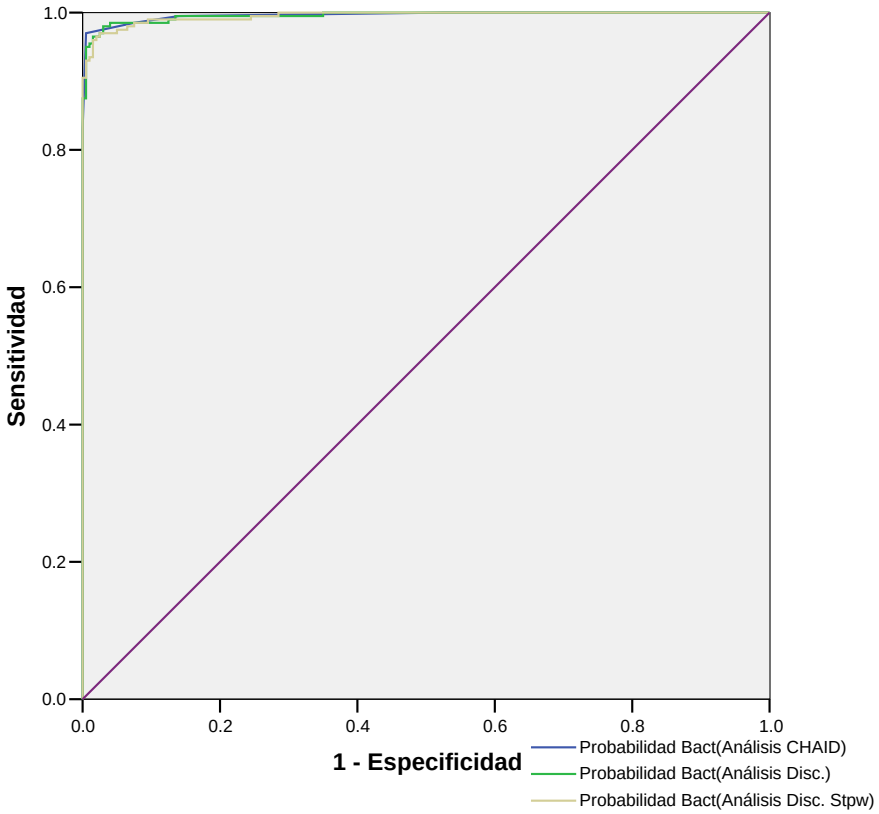


Figura 3.2.3.1.1 Curvas ROC obtenidas con los dos métodos de discriminante y con el método CHAID para bacterias.

Tabla 3.2.3.1.5 Parámetros calculados a partir de la matriz de confusión para evaluar el desempeño de los clasificadores utilizados.

Predicciones de los miembros del Grupo con Anl. Discriminante (Stpw).							
		Grupos	**Razón de TP**	**Razón de TN**	**Prec.**	**Exac.**	**% de Clasf.**
70 % base de datos extendida		**Archaea**	99.3	93.2	93.7	96.3	99.3
		Bacteria	93.2	99.3	99.3	96.3	93.2
Validación cruzada		**Archaea**	99.3	91.1	91.9	95.3	99.3
		Bacteria	91.1	99.3	99.3	95.3	91.1
Validación externa		**Archaea**	96.1	98.1	98.0	97.1	96.1
		Bacteria	98.1	96.1	96.4	97.1	98.1
Predicciones de los miembros del Grupo con Anl. Discriminante (Todas).							
70 % base de datos extendida		**Archaea**	100.0	93.8	94.3	96.9	100.0
		Bacteria	93.8	100.0	100.0	96.9	93.8
Validación cruzada		**Archaea**	99.3	90.4	91.4	94.9	99.3
		Bacteria	90.4	99.3	99.2	94.9	90.4
Validación externa		**Archaea**	98.0	98.1	98.0	98.1	98.0
		Bacteria	98.1	98.0	98.1	98.1	98.1
Predicciones de los miembros del Grupo con CHAID							
70 % base de datos extendida		**Archaea**	99.3	96.6	96.7	98.0	99.3
		Bacteria	96.6	99.3	99.3	98.0	96.6
Validación externa		**Archaea**	100.0	98.1	98.1	99.0	100.0
		Bacteria	98.1	100.0	100.0	99.0	98.1

3.2.4. Aminoácidos asociados con la clasificación taxonómica en vertebrados e invertebrados.

Cuando se aplica la técnica CHAID a la base curada con una validación cruzada se tienen los datos de la Tabla 3.2.4.1, donde podemos observar la correlación de los aminoácidos exceptuado la Cisteína la que tiene una significación mayor que 0.05, tenemos aquí el mejor porciento de clasificación presente en la Leucina, Tabla 3.2.4.2 y el que mejor significación presenta Ácido Aspártico, por lo que aparece en el nodo principal del árbol, Figura 3.2.4.1, donde además aparecen en nodos secundarios la Isolecina, Ácido

Glutámico y Triptófano aminoácidos con buena significación y buen porciento de clasificación. En la base de datos extendida con una validación del 70% de la muestra los resultados no son aceptables Tabla 3.2.4.3, pero fueron usados para la comparación con otro clasificador.

Tabla 3.2.4.1. Significación de los aminoácidos al ser utilizados como variables predictoras en la construcción de árboles de decisión y los porcientos de clasificación alcanzados.

AA	Sig.	%clasificación
Histidina	0.000313443	96,7
Tirosina	4.26129E-05	94
Glutamina	4.56215E-06	96
Fenilalanina	7.25884E-07	97
Serina	1.23944E-07	93,7
Alanina	2.88E-08	94,7
Glicina	2.22134E-09	96
Valina	1.14447E-09	96,3
Isoleucina	9.30623E-10	95
Metionina	2.24192E-15	95,7
Ácido Glutámico	6.51854E-18	96,3
Prolina	3.3356E-18	95
Lisina	1.04851E-26	95,7
Asparagina	6.24155E-27	97
Treonina	2.84273E-29	96,3
Leucina	1.56097E-35	98
Arginina	3.50044E-39	97,3
Triptófano	8.15877E-47	95,3
Ácido Aspártico	5.74625E-53	97

[a]Sig. Significación del estadígrafo de razón verosimilitud Chi-cuadrado. Por simplificación se ha utilizado el simbolismo del SPSS para la notación científica, es decir, por ejemplo, el símbolo E-05 significa 10^{-5}.

Tabla 3.2.4.2. Clasificación obtenida con método CHAID en la bases de datos curada con validación cruzada.

Muestras Observadas	Predicciones		
	vertebrados	invertebrados	Exactitud
vertebrados	193	7	96.5%
invertebrados	2	98	98.0%
% Total	65.0%	35.0%	97.0%

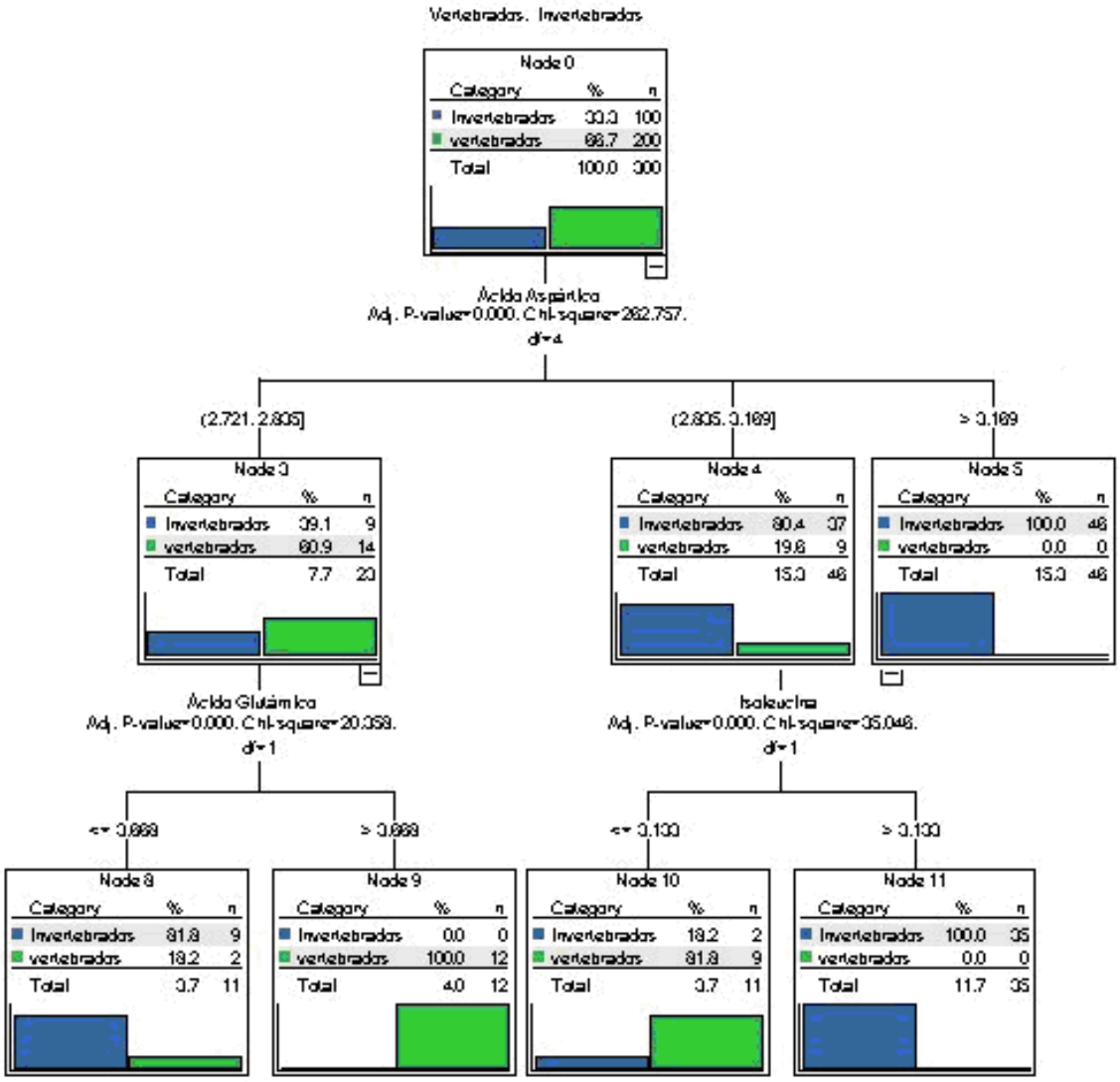


Figura 3.2.4.1A. Árbol de Aminoácidos asociados con los resultados en la base de datos curada con validación cruzada en las clasificaciones taxonómicas de vertebrados e invertebrados.

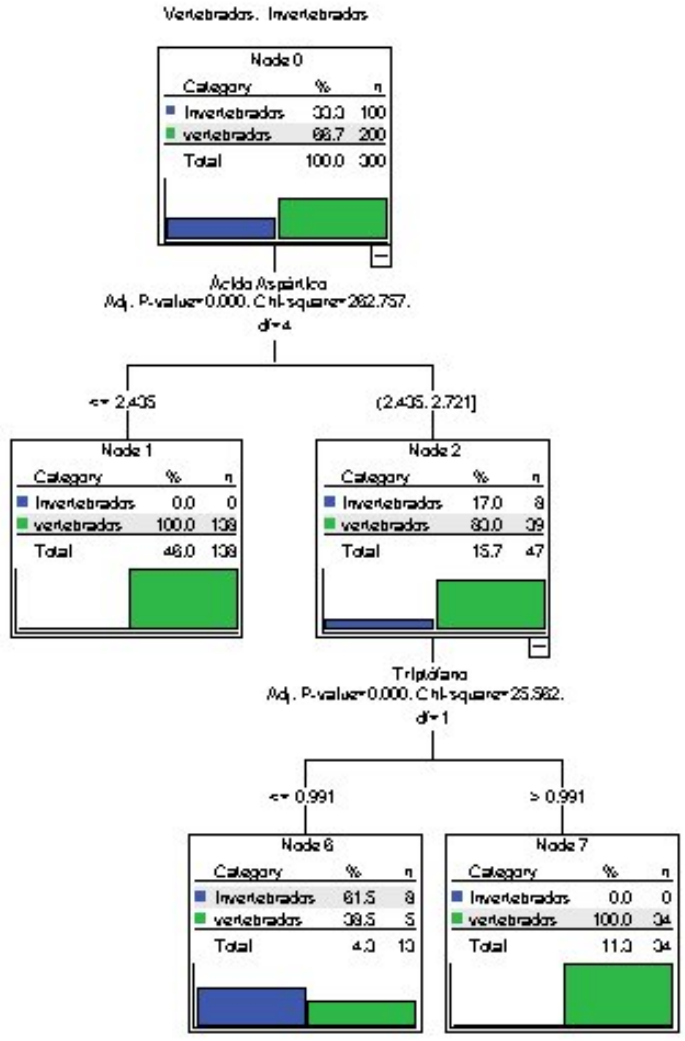


Figura 3.2.4.1.B Árbol de Aminoácidos asociados con los resultados en la base de datos curada con validación cruzada en las clasificaciones taxonómicas de vertebrados e invertebrados.

Tabla 3.2.4.3. Clasificación obtenida con método CHAID en la nueva base de datos extendida tomando aleatoriamente el 70% de la base como entrenamiento y el resto usado en validación externa.

		Predicciones		
Muestra	Observadas	vertebrados	invertebrados	Exactitud
Entrenamiento	vertebrados	64	8	88.9%
	invertebrados	9	67	88.2%
	% Total	49.3%	50.7%	88.5%
Validac.	vertebrados	22	6	78.6%
	invertebrados	7	17	70.8%
	% Total	55.8%	44.2%	75.0%

3.2.4.1. Análisis de Discriminante y la evaluación del desempeño de los clasificadores.

En esta taxa los resultados obtenidos con el CHAID, en la base extendida con una validación del 70% no fueron satisfactorios como fue discutido anteriormene. Con el método de Discriminante en la Tabla 3.2.4.1.1 se puede ver que en el caso del aminoácido Tirosina que no se incluye en el método Stepwise poseen correlación mayor que los demás incluídos, Tabla 3.2.4.1.2, el aminoácido Prolina que no se incluye para el caso donde se incluyen todos es el que presenta mayor valor de correlación.

Mientras, en la Tabla 3.2.4.1.3 se puede apreciar que la eficacia de las funciones discriminantes en la separación de los casos en grupos, expresada a través de las correlaciones canónicas, es similar para ambos procedimientos. Los valores de la Lambda de Wilk y la significación del test Chi-cuadrado indican que las capacidades discriminatorias de las funciones obtenidas por estos procedimientos son similares.

Tabla 3.2.4.1.1. Correlaciones de las variables discriminantes con las funciones discriminantes canónicas.

Aminoácido	Función discriminante
Prolina [a]	-0.276
Tirosina	-0.197
Ácido Aspártico	0.193
Treonina [a]	-0.181
Asparagina	0.166
Valina	0.124
Arginina	0.118
Serina	-0.113
Triptófano [a]	-0.112
Glicina	0.106
Leucina [a]	-0.095
Fenilalanina	0.061
Glutamina [a]	-0.051
Cisteína [a]	-0.049
Lisina [a]	0.039
Histidina [a]	0.031
Metionina	-0.008
Isoleucina [a]	-0.006
Alanina	0.003
Ácido Glutámico	-0.002

Tabla 3.2.4.1.2. Funciones discriminantes canónicas obtenidas con la introducción de todos los aminoácidos que satisfacen el test de tolerancia y con el método Stepwise.

Aminoácidos	Función discriminante	
	Todas	Stepwise
Alanina	3.33494748	1.13469444
Cisteína	1.81042056	-
Ácido Aspártico	5.96872564	4.72640862
Ácido Glutámico	0.29380426	-1.42722167
Fenilalanina	3.34144515	1.54141778
Glicina	0.81918569	0.63452814
Histidina	0.95372764	-
Isoleucina	2.96069855	-
Lisina	1.74121503	-
Leucina	1.04259695	-
Metionina	5.07536269	3.45884755
Asparagina	3.99586851	2.83456292
Prolina	2.01832702	-
Glutamina	1.89329619	-
Arginina	4.79244112	2.67161885
Serina	3.28271896	1.41811029
Treonina	1.25945216	-
Valina	3.05596037	1.45824349
Tirosina	-	-2.0230344
Triptófano	2.51040997	-
(Constant)	148.707461	-44.24496

Los resultados de clasificación global para los métodos de obtención de las funciones discriminantes y para el método CHAID se observan en las curvas ROC obtenidas (Figura 3.2.4.1.1) y en la Tabla 3.2.4.1.4 de área bajo la curva donde la superioridad del Discriminante queda clara en los datos de intervalos de confianza asintóticos para 95%, quedando totalmente incluído el intervalo obtenido del CHAID en el obtenido del Discriminante.

Tabla 3.2.4.1.3. Eficacia de las funciones discriminantes a través de las correlaciones canónicas y los valores de la Lambda de Wilk.

Función	Valor principal	% de Varianza	% Var. Acum.	Corr. Canónica	Función	Lambda de Wilks	Chi cuadrado	g.l.	Sig.
Stepwise									
1	6.659	100	100	0.932	1	0.131	286.037	11	0.000
Todas las variables									
1	7.146	100	100	0.937	1	0.123	286.316	19	0.000

Tabla 3.2.4.1.4. Resultado del área bajo la curva en los tres métodos utilizados.

Resultados del Análisis	Área	Error Estándar	Sig. Asintótica	Intervalo de confianza asintótico para el 95%	
				Límite superior	Límite inferior
Probabilidad Vertebrado (Análisis Disc. Stepwise)	1.000	0.001	0.000	0.998	1.001
Probabilidad Vertebrado (Análisis CHAID)	0.951	0.014	0.000	0.924	0.978
Probabilidad Vertebrado (Análisis Disc. Todas)	1.000	0.000	0.000	1.000	1.000

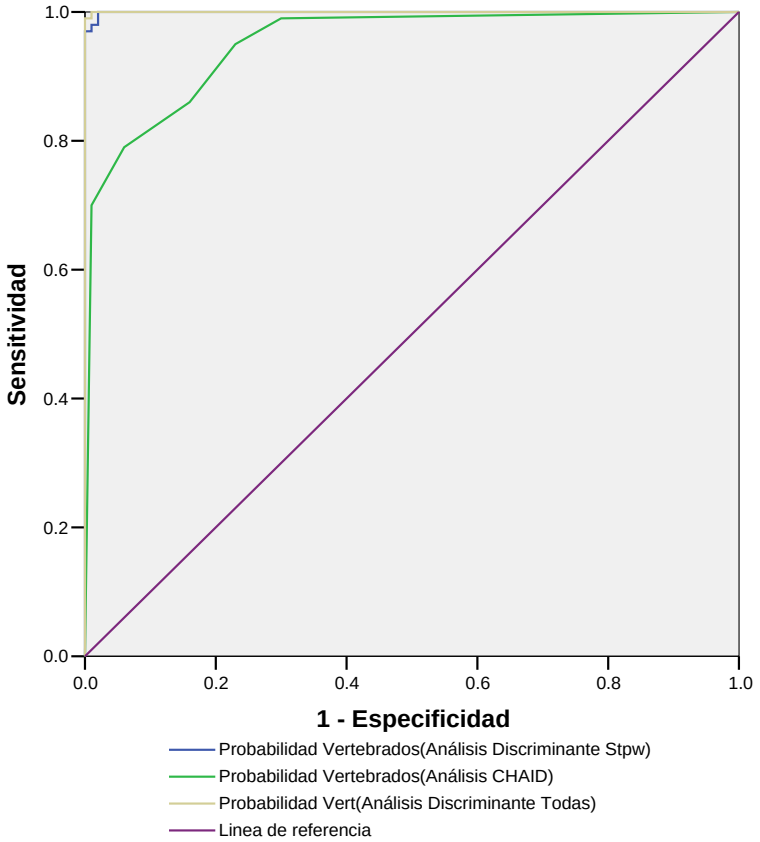


Figura 3.2.4.1.1 Curvas ROC obtenidas con los dos métodos de discriminante y con el método CHAID para vertebrados.

De los parámetros de la matriz de confusión Tabla 3.2.4.1.5, se muestra también que la diferencia rádica en el hecho que el método de Discriminante muestra valores superiores en todos los parámetros.

Tabla 3.2.4.1.5 Parámetros calculados a partir de la matriz de confusión para evaluar el desempeño de los clasificadores utilizados.

	Grupos	Razón de TP	Razón de TN	Prec.	Exac.	% de Clasf.
Predicciones de los miembros del Grupo con Anl. Discriminante (Stpw).						
70 % base de datos extendida	Vert	98.6	98.7	98.6	98.6	98.6
	Invert	98.7	98.6	98.7	98.6	98.7
Validación cruzada	Vert	98.6	98.7	98.6	98.6	98.6
	Invert	98.7	98.6	98.7	98.6	98.7
Validación externa	Vert	100.0	95.8	96.6	63.8	100.0
	Invert	95.8	100.0	45.1	63.8	95.8
Predicciones de los miembros del Grupo con Anl. Discriminante (Todas).						
70 % base de datos extendida	Vert	98.6	98.7	98.6	98.6	98.6
	Invert	98.7	98.6	98.7	98.6	98.7
Validación cruzada	Vert	97.2	98.7	98.6	98.0	97.2
	Invert	98.7	97.2	97.4	98.0	98.7
Validación externa	Vert	100.0	100.0	100.0	100.0	100.0
	Invert	100.0	100.0	100.0	100.0	100.0
Predicciones de los miembros del Grupo con CHAID						
70 % base de datos extendida	Vert	88.9	88.2	87.7	88.5	88.9
	Invert	88.2	88.9	89.3	88.5	88.2
Validación externa	Vert	78.6	70.8	75.9	75.0	78.6
	Invert	70.8	78.6	73.9	75.0	70.8

3.2.5. Aminoácidos asociados con la clasificación taxonómica en vertebrados no mamíferos y mamíferos.

El interés biológico en el estudio de esta taxa esta dado por el hecho que ella representa a dos grupos de organismos que durante el proceso evolutivo ocurre su separación en un determinado momento por lo que sugiere que compartan un número importante de caracteres y que para su diferenciación sea importante contar con otro criterio como el que nos proponemos verificar en esta sección con las pruebas estadísticas realizadas. Cuando se aplica la técnica CHAID, Tabla 3.2.5.1 podemos observar que el aminoácido que tiene mejor porciento de clasificación es aquel que mayor signifcación posee, la Metionina.

Tabla 3.2.5.1. Significación de los aminoácidos al ser utilizados como variables predictoras en la construcción de árboles de decisión y los porcientos de clasificación alcanzados.

AA	Sig.	%clasificación
Glicina	0.031889439	88
Serina	0.006475131	89,5
Tirosina	0.000723822	90
Lisina	0.000384934	89,5
Triptófano	0.000366596	87
Fenilalanina	0.000173892	87
Alanina	4.58E-05	93
Histidina	2.33556E-05	90,5
Cisteína	1.52071E-05	92
Ácido Aspártico	9.17068E-09	90
Glutamina	3.65863E-11	89
Leucina	1.81915E-12	90,5
Arginina	1.97404E-13	93,5
Prolina	1.28088E-13	93,5
Isoleucina	3.88711E-14	91,5
Ácido Glutámico	2.8872E-15	91
Treonina	5.19733E-16	92,5
Valina	4.30663E-16	94
Asparagina	3.10522E-16	92
Metionina	1.96905E-21	95

[a]Sig. Significación del estadígrafo de razón verosimilitud Chi-cuadrado. Por simplificación se ha utilizado el simbolismo del SPSS para la notación científica, es decir, por ejemplo, el símbolo E-05 significa 10^{-5}.

Tabla 3.2.5.2. Clasificación obtenida con método CHAID en la bases de datos curada con validación cruzada.

Muestras Observadas	Predicciones		
	vertebrados no mamiferos	mamiferos	Exactitud
vertebrados no mamiferos	47	3	94.0%
mamiferos	7	143	95.3%
% Total	27.0%	73.0%	95.0%

En la Tabla 3.2.5.2 se muestran los resultados de una validación cruzada en la base de datos curada y en la Figura 3.2.5.1 muestra el árbol que, además de tener en su nodo principal el aminoácido de mayor significación, intervienen otros en los nodos secundarios, Prolina y Asparagina, que presentan alta significación por lo que esta altamente correlacionados.

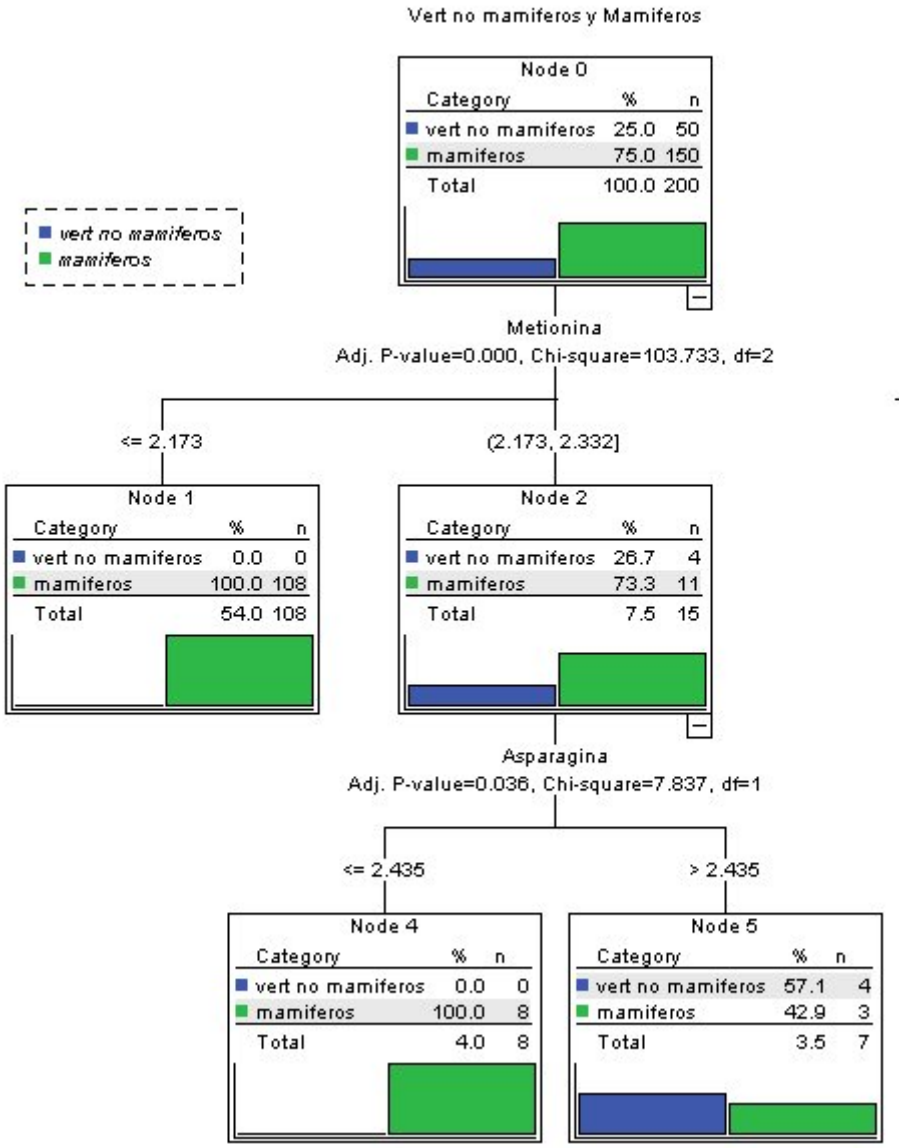


Figura 3.2.5.1 Sección A del árbol de aminoácidos asociados con los resultados en la base de datos curada con validación cruzada en las clasificaciones taxonómicas de vertebrados no mamiferos y mamiferos.

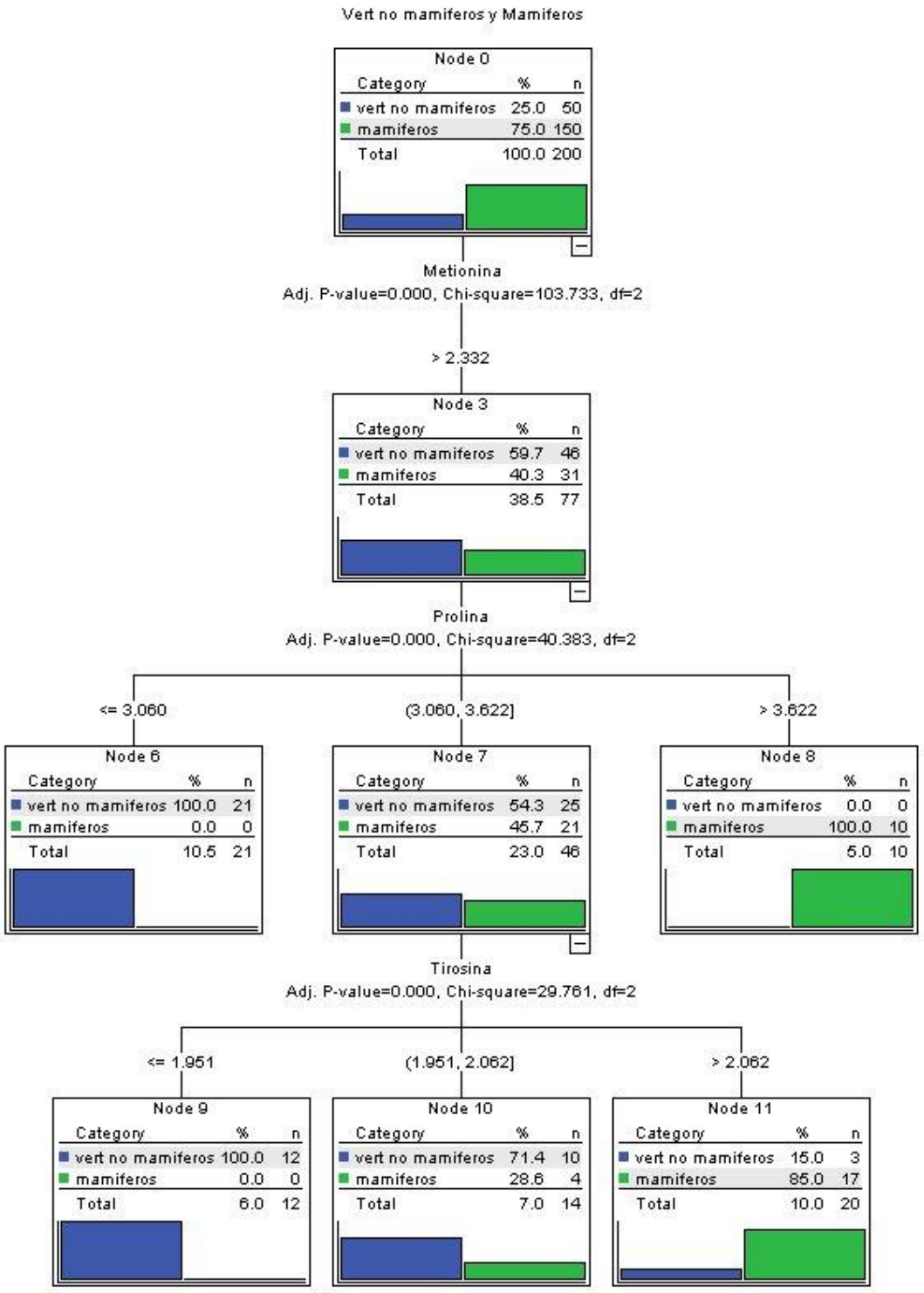


Figura 3.2.5.1 Sección B del árbol de aminoácidos asociados con los resultados en la base de datos curada con validación cruzada en las clasificaciones taxonómicas de vertebrados no mamiferos y mamiferos.

Tabla 3.2.5.3 Clasificación obtenida con método CHAID en la nueva base de datos extendida tomando aleatoriamente el 70% de la base como entrenamiento y el resto usado en validación externa.

		Predicciones		
muestra	Observaciones	Mamiferos	vertNoMamif	Exactitud
entrena miento	Mamiferos	72	0	100.0%
	vertNoMamif	8	68	89.5%
	%total	54.1%	45.9%	94.6%
praueb	Mamiferos	27	1	96.4%
	vertNoMamif	2	22	91.7%
	%total	55.8%	44.2%	94.2%

Los resultados son corroborados con una base extendida donde los porcientos de clasificación son aceptables. En la Tabla 3.2.5.3 se puede apreciar que tanto en la base de entrenamiento (70% de la base) como en la base externa (resto de la base) se alcanza un 94%.

3.2.5.1. Análisis de Discriminante y la evaluación del desempeño de los clasificadores.

Con el análisis de discriminante pudimos comprobar que toods los aminoácidos están asociados con la clasificación de los vectores NEC_k en estos dos grupos de organismos. En la Tabla 3.2.5.1.1 se puede ver que todos poseen correlaciones altas de las variables con la función Discriminante. En la Tabla 3.2.5.1.2 se presentan las funciones discriminantes obtenidas por el método Stepwise minimizando la Lambda de Wilk y sin aplicar este método, aquí observamos que el aminoácido Tirosina no esta presente en ninguno de los dos métodos y que solo nueve aminoácidos estan presentes en el método Stepwise.

Mientras, en la Tabla 3.2.5.1.3 se puede apreciar que la eficacia de las funciones discriminantes en la separación de los casos en grupos, expresada a través de las correlaciones canónicas, es similar para ambos procedimientos. Además, los valores de la Lambda de Wilk y la significación del test Chi-cuadrado indican que las capacidades discriminatorias de las funciones obtenidas por estos procedimientos son similares, indicando el buen desempeño de las funciones discriminantes.

En las curvas ROC obtenidas (Figura 3.2.5.1.1) se muestran con claridad los tres métodos aplicados, sin embargo los resultados de clasificación global no son estadísticamente diferentes para los métodos de obtención de las funciones discriminantes y para el método CHAID, Tabla 3.2.5.1.4, en la que se muestra que los intervalos de confianza asintóticos para 95% de confianza de las áreas bajo la curva ROC se solapan. Cuando se utilizan los parámetros derivados de la matríz de confusión para evaluar el desempeño de estos clasificadores, nos sugieren que las diferencias entre los clasificadores son mínimas en la Tabla 3.2.5.1.5 se muestran los valores de los parámetros mencionados.

Tabla 3.2.5.1.1. Correlaciones de las variables discriminantes con las funciones discriminantes canónicas.

Aminoácido	Función Discriminante
Valina	0.601773568
Asparagina	-0.542521612
Isoleucina	-0.532015422
Ácido Glutámico	0.466758664
Leucina	-0.461167619
Glutamina	0.454836113
Arginina	0.442923251
Treonina	-0.416294569
Metionina	-0.403460246
Ácido Aspártico	0.356307615
Cisteína	0.25158214
Prolina	0.242783766
Fenilalanina	-0.239970835
Lisina	0.204346612
Glicina	0.199265288
Triptófano	-0.162943411
Alanina	-0.136532651
Tirosina	0.108252116
Histidina	0.078358099
Serina	-0.024180644

Tabla 3.2.5.1.2. Funciones discriminantes canónicas obtenidas con la introducción de todos los aminoácidos que satisfacen el test de tolerancia y con el método Stepwise.

Aminoácidos	Función discriminante	
	Todas	Stepwise
Alanina	-1.908191	-1.922293
Cisteína	0.1611483	-
Ácido Aspártico	-1.07872	-
Ácido Glutámico	0.5832623	-
Fenilalanina	-1.291696	-0.949225
Glicina	-0.42828	-0.665337
Histidina	-0.052023	-
Isoleucina	0.9644433	0.8220759
Lisina	0.3764664	-
Leucina	0.234707	-
Metionina	-0.770816	-
Asparagina	-1.404127	-1.590907
Prolina	1.9039964	2.2351536
Glutamina	1.3136492	1.7373429
Arginina	-0.791883	-
Serina	0.5240509	-
Treonina	-0.775059	-
Valina	2.7165784	2.8706175
Triptófano	3.5642005	2.8976621
(Constant)	-6.41049	-9.487415

Tabla 3.2.5.1.3. Eficacia de las funciones discriminantes a través de las correlaciones canónicas y los valores de la Lambda de Wilk.

Función	Valor principal	% de Varianza	% Var. Acum.	Corr. Canónica	Función	Lambda de Wilks	Chi cuadrado	g.l.	Sig.
Stepwise									
1	3.481	100	100	0.881	1	0.223	212.232	9	0.000
Todas las variables									
1	3.849	100	100	0.891	1	0.206	215.502	19	0.000

Tabla 3.2.5.1.4. Resultado del área bajo la curva en los tres métodos utilizados.

Resultados del Análisis	Área	Error Estándar	Sig. Asintótica	Intervalo de confianza asintótico para el 95%	
				Límite superior	Límite inferior
Probabilidad Mamíferos (Análisis CHAID)	0.974	0.010	0.000	0.953	0.994
Probabilidad Mamíferos (Análisis Discriminante)	0.999	0.001	0.000	0.996	1.001
Probabilidad Mamíferos (Análisis Disc. Stepwise)	0.996	0.002	0.000	0.991	1.000

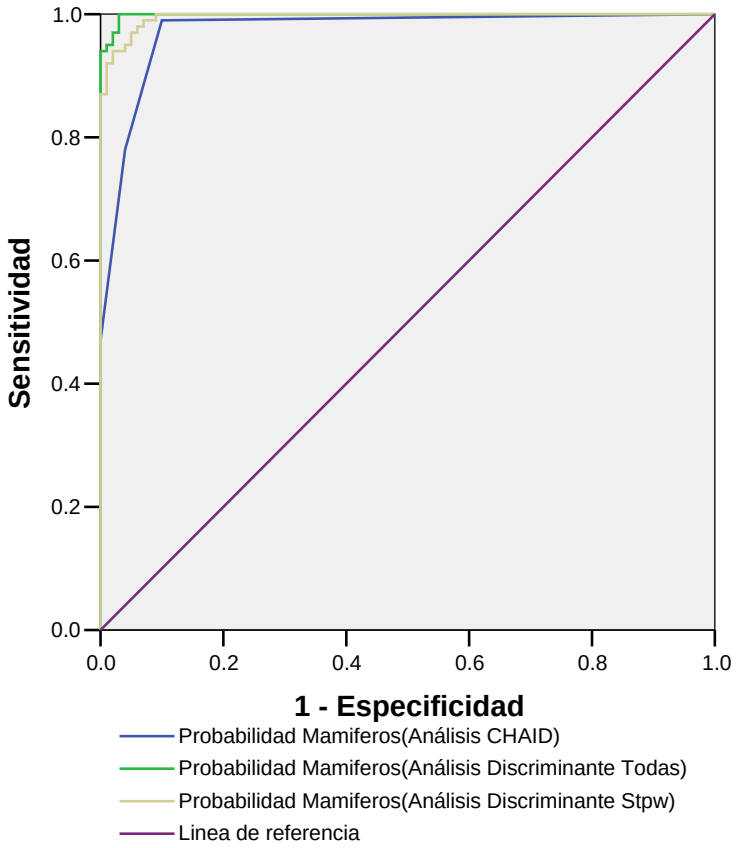


Figura 3.2.5.1.1 Curvas ROC obtenidas con los dos métodos de discriminante y con el método CHAID para mamíferos.

Tabla 3.2.5.1.5 Parámetros calculados a partir de la matriz de confusión para evaluar el desempeño de los clasificadores utilizados.

Predicciones de los miembros del Grupo con Anl. Discriminante (Stpw).							
		Grupos	**Razón de TP**	**Razón TN**	**Prec.**	**Exac.**	**% de Clasf.**
70 % base de datos extendida		Mamiferos	95.8	96.1	95.8	95.9	95.8
		VertNoMamif	96.1	95.8	96.1	95.8	96.1
Validación cruzada		Mamiferos	94.4	96.1	95.8	95.3	94.4
		VertNoMamif	96.1	94.4	94.8	95.3	96.1
Validación externa		Mamiferos	96.4	91.7	93.1	94.2	96.4
		VertNoMamif	91.7	96.4	95.7	94.2	91.7
Predicciones de los miembros del Grupo con Anl. Discriminante (Todas).							
70 % base de datos extendida		Mamiferos	97.2	93.1	97.2	97.3	97.2
		VertNoMamif	93.1	97.2	97.4	97.3	97.4
Validación cruzada		Mamiferos	93.1	94.7	94.4	93.9	93.1
		VertNoMamif	94.7	93.1	93.5	93.9	94.7
Validación externa		Mamiferos	96.4	95.8	96.4	96.2	96.4
		VertNoMamif	95.8	96.4	95.8	96.2	95.8
Predicciones de los miembros del Grupo con CHAID							
70 % base de datos extendida		Mamiferos	100.0	89.5	90.0	94.6	100.0
		VertNoMamif	89.5	100.0	100.0	94.6	89.5
Validación externa		Mamiferos	96.4	91.7	93.1	94.2	96.4
		VertNoMamif	91.7	96.4	95.7	94.2	91.7

3.2.6. Aminoácidos asociados con la clasificación taxonómica en primates y homo sapiens.

Por las especies que involucra esta taxa se hace particularmente interesante el análisis si tenemos en cuenta que, además de todas las peculiaridades de las proteínas vistas en el Capítulo 2, se puede agregar que los Homo Sapiens y los primates pertenecientes ambos al orden primate, clase mamíferos, la similitud entre sus DNA llega a ser en algunas especies de hasta un 98,5 % (ejemplo homo sapiens y chimpancé). Como se puede observar en la Tabla 3.2.6.1, no todos los aminoácidos alcanzan una buena significación, en ese caso están la Glutamina, la Cisteína, la Fenilalanina, la Asparagina, la Arginina y la Serina.

Mientras la Metionina, la Tirosina, la Glicina y la Leucina logran un 97 % de clasificación, siendo la Leucina el aminoácido que posee la mayor significación, este resultado se obtiene con la base de datos curada ver Tabla 3.2.6.2 y Figura 3.2.6.1.

Es de esperar desde el punto de vista Biológico que cuando el análisis se realiza en una base de datos extendida con una validación del 70% de la muestra los porcientos de clasificación no sean tan buenos ver Tabla 3.2.6.3, sin embargo nuestro propósito es verificar que el uso de los vectores NEC_K de las frecuencias de probabilidades de los aminoácidos en cadenas de proteínas para esta taxa logra una diferenciación clara entre las dos especies involucradas, a pesar de su similitud en este orden.

Tabla 3.2.6.1. Significación de los aminoácidos al ser utilizados como variables predictoras en la construcción de árboles de decisión y los porcientos de clasificación alcanzados.

AA	Sig.	%clasificación
Metionina	0.043322937	97
Valina	0.006443999	92
Prolina	0.005132947	94
Alanina	1.28E-03	96
Tirosina	0.000115772	97
Histidina	2.05213E-05	93
Isoleucina	6.71297E-09	94
Glicina	2.52018E-09	97
Lisina	3.44484E-10	95
Treonina	8.23E-13	93
Ácido Aspártico	3.62791E-13	93
Triptófano	3.26E-13	92
Ácido Glutámico	7.29929E-14	94
Leucina	1.09352E-21	97

[a]Sig. Significación del estadígrafo de razón verosimilitud Chi-cuadrado. Por simplificación se ha utilizado el simbolismo del SPSS para la notación científica, es decir, por ejemplo, el símbolo E-05 significa 10^{-5}.

Tabla 3.2.6.2. Clasificación obtenida con método CHAID en la bases de datos curada con validación cruzada.

Muestras Observadas	Predicciones		
	primates	homo sapiens	Exactitud
primates	50	0	100.0%
homo sapiens	3	47	94.0%
% Total	53.0%	47.0%	97.0%

Tabla 3.2.6.3 Clasificación obtenida con método CHAID en la nueva base de datos extendida tomando aleatoriamente el 70% de la base como entrenamiento y el resto usado en validación externa.

		Predicciones		
muestra	observada	primate	homoSapiens	Exactitud
entrenamiento	primate	36	3	92.3%
	homoSapiens	4	33	89.2%
	% Total	52.6%	47.4%	90.8%
valid	primate	7	4	63.6%
	homoSapiens	2	11	84.6%
	% Total	37.5%	62.5%	75.0%

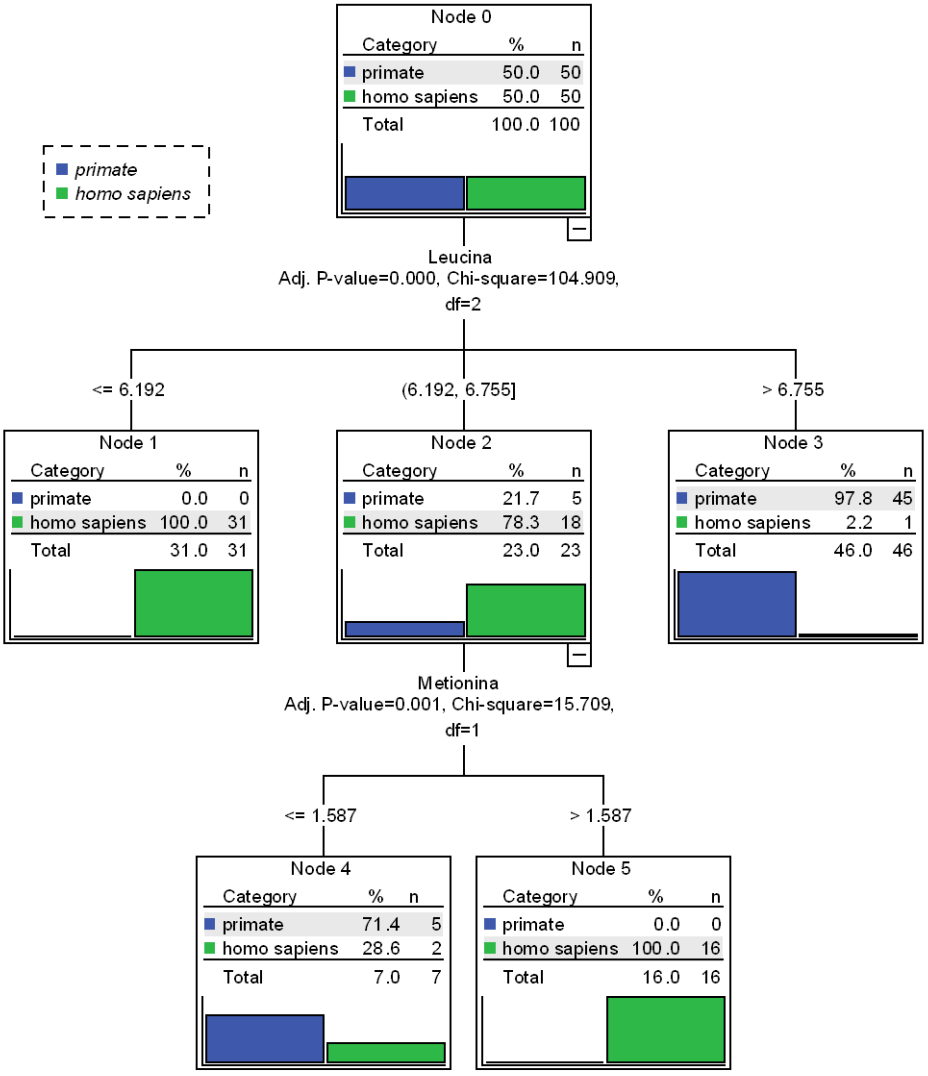


Figura 3.2.6.1 Árbol de aminoácidos asociados con los resultados en la base de datos curada con validación cruzada en las clasificaciones taxonómicas de primates y homo sapiens.

3.2.6.1. Análisis de Discriminante y la evaluación del desempeño de los clasificadores.

El análisis de discriminante realizado en esta taxa se corroboró el resultado, previamente obtenido con el CHAID, de que todos los aminoácidos están asociados con la clasificación de los vectores NEC_k en estos dos reinos. En la Tabla 3.2.6.1.1 se puede ver que, incluso, aminoácidos como la Tirosina, el cual no está incluídos en las combinaciones lineales de las funciones discriminantes en los dos métodos aplicados, posee una correlación que no es la mejor, pero si mayor que la que tienen la mayoria de los que están incluídos ver Tabla 3.2.6.1.2. En la Tabla 3.2.6.1.1, se observa como los aminoácidos Lisina, Ácido Aspártico, Triptófano y Ácido Glutámico poseen los mayores coeficientes de correlación absolutos y altamente significativos y además todos se incluyen en las funciones discriminantes aplicadas (Tabla 3.2.6.1.2). Mientras, en la Tabla 3.2.6.1.3 se puede apreciar que los valores de la Lambda de Wilk y la significación del test Chi-cuadrado, indican que las capacidades discriminatorias de las funciones obtenidas por estos procedimientos son similares. En particular, para todas las funciones los valores de estos parámetros son altos, indicando el buen desempeño de las funciones discriminantes. Además los indicadores de la correlación canónica indican la eficacia de la funciones por los valores próximos obtenidos en ambos métodos.

Cuando se evalua el desempeño de los métodos en la Figura 3.2.6.1.1, con la construcción de las curvas ROC y en la Tabla 3.2.6.1.4, podemos observar que con el método Discriminante en sus dos variantes se obtiene 100% de clasificación no siendo así con el CHAID que se obtiene un 95%, sin embargo no consideramos que estas diferencias sean estadísticamente significativas si tenemos en cuenta las carácterísticas del taxa con que se trabaja.

Cuando se calculan los parámetros a partir de las matrices de confusión se observa en la Tabla 3.2.6.1.5 que las diferencias son más pronunciadas entre los clasificadores pues mientras que para los análisis de Discriminante los parámetros están por encima de un 95% para el CHAID y en particular la clasificación de Homo Sapiens presenta porcientos no aceptables.

Tabla 3.2.6.1.1. Correlaciones de las variables discriminantes con las funciones discriminantes canónicas.

Aminoácido	**Función discriminante**
Lisina	0.4079589
Ácido Aspártico	0.366458
Triptófano	-0.327464
Ácido Glutámico	0.3153944
Leucina	-0.292727
Tirosina	-0.251821
Glicina	0.2073298
Histidina	0.2034503
Treonina	-0.199123
Prolina	0.1655295
Valina	-0.116912
Cisteína	-0.112406
Isoleucina	-0.097109
Arginina	0.0765084
Alanina	-0.072726
Serina	-0.07265
Fenilalanina	-0.069234
Metionina	-0.061917
Glutamina	0.0590462
Asparagina	0.0554707

Tabla 3.2.6.1.2. Funciones discriminantes canónicas obtenidas con la introducción de todos los aminoácidos que satisfacen el test de tolerancia y con el método Stepwise.

Aminoácidos	Función discriminante	
	Todas	Stepwise
Alanina	-0.120741	-
Cisteína	-0.112789	-
Ácido Aspártico	1.2602754	1.3495373
Ácido Glutámico	1.6410213	1.3228442
Fenilalanina	1.0233814	1.0276529
Glicina	0.9222785	1.1378398
Histidina	1.8013634	2.0428226
Isoleucina	0.2280033	-
Lisina	0.8002481	1.1375742
Leucina	-0.844817	-0.747004
Metionina	2.740958	3.3493038
Asparagina	-0.54982	-
Prolina	2.0142909	3.3493038
Glutamina	-0.023842	-
Arginina	-0.991585	-
Serina	0.1109863	-
Treonina	-0.717032	-
Valina	0.2848558	-
Triptófano	-2.760603	-3.129541
(Constant)	-18.40188	-26.26685

Tabla 3.2.6.1.3. Eficacia de las funciones discriminantes a través de las correlaciones canónicas y los valores de la Lambda de Wilk.

Función	Valor principal	% de Varianza	% Var. Acum.	Corr. Canónica	Función	Lambda de Wilks	Chi cuadrado	g.l.	Sig.
Stepwise									
1	6.352	100	100	0.930	1	0.136	191.516	10	0.00
Todas las variables									
1	6.942	100	100	0.935	1	0.126	189.601	19	0.00

Tabla 3.2.6.1.4. Resultado del área bajo la curva en los tres métodos utilizados.

Resultados del Análisis	Área	Error Estándar	Sig. Asintótica	Intervalo de confianza asintótico para el 95%	
				Límite superior	Límite inferior
Probabilidad Homo Sapiens (Análisis Disc. Stepwise)	1.000	0.000	0.000	1.000	1.000
Probabilidad Homo Sapiens (Análisis CHAID)	0.949	0.020	0.000	0.911	0.988
Probabilidad Homo Sapiens (Análisis Discriminante)	1.000	0.000	0.000	1.000	1.000

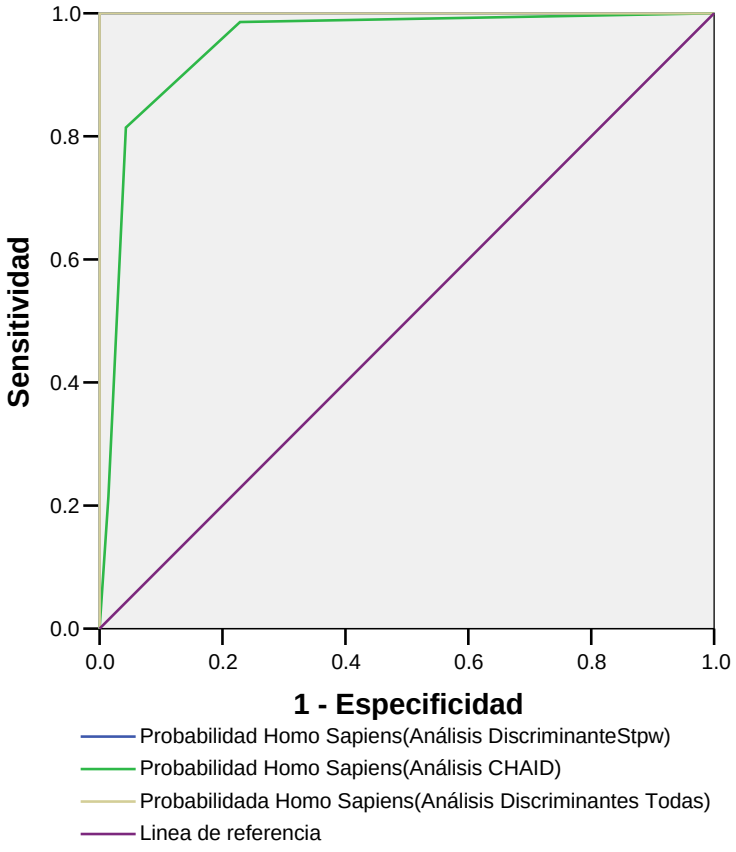


Figura 3.2.6.1.1 Curvas ROC obtenidas con los dos métodos de discriminante y con el método CHAID para Homo Sapiens.

Tabla 3.2.6.1.5 Parámetros calculados a partir de la matriz de confusión para evaluar el desempeño de los clasificadores utilizados.

Predicciones de los miembros del Grupo con Anl. Discriminante (Stpw).							
		Grupos	**Razón de TP**	**Razón TN**	**Prec.**	**Exac.**	**% de Clasf.**
70 % base de datos extendida		Primates	98.2	100.0	100.0	99.0	98.2
		HomoS	100.0	98.2	97.9	99.0	100.0
Validación cruzada		Primates	98.2	100.0	100.0	99.0	98.2
		HomoS	100.0	98.2	97.9	99.0	100.0
Validación externa		Primates	100.0	100.0	100.0	100.0	100.0
		HomoS	100.0	100.0	100.0	100.0	100.0
Predicciones de los miembros del Grupo con Anl. Discriminante (Todas).							
70 % base de datos extendida		Primates	98.2	100.0	100.0	99.0	98.2
		HomoS	100.0	98.2	97.9	99.0	100.0
Validación cruzada		Primates	96.4	97.9	98.2	97.1	96.4
		HomoS	97.9	96.4	95.8	97.1	97.9
Validación externa		Primates	100.0	100.0	100.0	100.0	100.0
		HomoS	100.0	100.0	100.0	100.0	100.0
Predicciones de los miembros del Grupo con CHAID							
70 % base de datos extendida		Primate	96.4	89.4	91.5	93.2	96.4
		HomoS	89.4	95.5	95.5	93.2	89.4
Validación externa		Primate	92.9	65.2	61.9	75.7	92.9
		HomoS	65.2	93.8	93.8	75.7	65.2

3.3. Construcción de árboles de clasificación mediante el método CHAID atendiendo a las frecuencias del uso de codones de los aminoácidos en los genes.

Al pasar de una secuencia de codones a la correspondiente secuencia de aminoácidos se pierde información debido a la degeneración del código genético (ver sección 1.2). Por tal motivo, pudiera pensarse que ocurra un cambio en los vectores NEC_k tal que afecte la clasificación de los taxa. Luego, se hace necesaria la verificación de la hipótesis de investigación partiendo de secuencias de genes. En nuestro caso, como se explicó en el capítulo 2, se utilizó la información recopilada en la base de datos de uso de codones.

Los análisis se realizaron utilizando como entrenamiento el 70% de las bases de datos de los taxa construidas. Se realizó validación cruzada con la base de entrenamiento y una validación externa con el 30% restante.

3.3.1. Aminoácidos asociados mediante el uso de codones con las clasificaciones taxonómicas en organismos vivos.

En el análisis realizado con todos los taxa se obtienen bajos porcientos de clasificación. Sin embargo, al igual que el resultado obtenido con las bases de secuencias de proteínas, se verificó que todos los aminoácidos están asociados de manera altamente significativa con la clasificación biológica. En particular, los resultados obtenidos con CHAID se resumen en la Tabla 3.3.1.1, donde podemos observar que el aminoácido con mayor significación es el que produce mayor porciento de clasificación aunque no sea bueno, mientras con el análisis de Discriminante en las Tablas 3.3.1.2 y 3.3.1.3, por ejemplo la Tirosina, el cual no está incluído en las combinaciones lineales de las funciones discriminantes para el caso en que intervienen todos los aminoácidos si esta presente cuando se aplica el método de Stepwies, además presenta correlaciones altas con la funciones discriminantes.

Aunque, al igual que con la base de aminoácidos, los porcientos de clasificación correcta obtenidos con el análisis de Discriminante son mejores que con el método CHAID, los resultados sugieren que es posible alcanzar una mayor significación estadística en la diferenciación de los taxa si se analizan por separados grupos de taxa atendiendo a criterios de interés biológicos-evolutivos.

Tabla 3.3.1.1. Significación de los aminoácidos al ser utilizados como variables predictoras en la construcción de árboles de decisión y los porcientos de clasificación alcanzados.

AA	Sig.	%clasificación
Fenilalanina	1.11011E-19	62,9
Metionina	4.56265E-29	63,6
Arginina	5.26252E-30	64,2
Triptófano	7.63E-36	66,9
Prolina	7.93253E-41	67,8
Tirosina	1.96584E-42	68
Leucina	2.69777E-47	69,8
Ácido Glutámico	1.27706E-48	66
Histidina	1.83529E-49	67,3
Ácido Aspártico	1.11934E-51	68,4
Glicina	3.50217E-52	66,2
Lisina	2.17215E-54	68,7
Asparagina	3.46256E-61	67,3
Treonina	2.64E-61	66,4
Isoleucina	1.55462E-61	64
Cisteína	4.30445E-70	66
Glutamina	1.23215E-77	65,1
Valina	1.52126E-78	70
Alanina	4.39E-93	70,2
Serina	3.2243E-101	72,4

[a]Sig. Significación del estadígrafo de razón verosimilitud Chi-cuadrado. Por simplificación se ha utilizado el simbolismo del SPSS para la notación científica, es decir, por ejemplo, el símbolo E-05 significa 10^{-5}.

Tabla 3.3.1.2. Correlaciones de las variables discriminantes con las funciones discriminantes canónicas.

Amino ácidos	Función Discriminante							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Serina	-0.666	0.174	0.039	-0.024	-0.180	0.106	0.118	0.234
Alanina	0.284	-0.472	-0.194	-0.313	-0.041	0.030	-0.083	0.384
Leucina	0.034	0.029	-0.504	-0.120	-0.026	0.013	0.348	-0.369
Asparagina	-0.162	-0.020	0.461	0.329	-0.069	0.212	-0.371	-0.206
Triptófano	0.004	-0.142	-0.416	0.084	-0.337	-0.191	-0.079	-0.144
Ácido Aspártico	0.060	-0.227	0.396	-0.174	0.240	0.167	0.095	-0.176
Tirosina	0.051	0.159	0.137	0.410	-0.204	0.240	0.008	-0.247
Isoleucina	0.231	0.390	0.178	0.103	-0.394	0.030	0.107	-0.361
Prolina	-0.107	-0.088	-0.115	-0.126	0.373	0.166	-0.153	0.292
Ácido Glutámico	0.180	0.293	0.170	-0.162	0.345	-0.166	-0.163	0.155
Treonina	-0.096	-0.285	-0.114	0.312	0.144	-0.407	0.106	0.143
Valina	0.361	0.011	-0.049	-0.094	-0.344	0.150	0.505	-0.058
Metionina	0.040	0.096	0.247	0.065	-0.047	-0.318	0.488	-0.063
Lisina	0.060	0.273	0.281	-0.037	0.010	0.097	-0.292	-0.253
Glicina	0.089	-0.132	0.011	-0.138	0.063	-0.187	-0.240	-0.145
Glutamina	-0.317	-0.122	-0.208	0.399	0.334	-0.004	-0.068	0.459
Arginina	0.064	-0.121	-0.104	-0.159	0.065	-0.279	-0.032	0.435
Fenilalanina	0.029	0.156	-0.005	0.146	-0.303	0.144	0.117	-0.404
Cisteína	-0.191	0.108	-0.027	0.170	-0.049	0.205	0.143	-0.276
Histidina	-0.147	-0.143	-0.022	0.162	0.038	-0.103	0.027	0.273

Tabla 3.3.1.3. Funciones discriminantes canónicas obtenidas con la introducción de todos los aminoácidos que satisfacen el test de tolerancia y con el método stepwise.

Amino ácidos	1		2		3		4		5		6		7		8	
	Todas	Stpws	Todas	Stpws	Todas	Stpws	Todas	Stpws	Todas	Stpws	Todas	Stpws	Todas	Stpws	Todas	Stpws
Ala	0.632	0.630	1.990	-0.905	0.146	0.177	1.449	-0.037	0.984	-0.304	0.398	1.037	0.434	0.385	0.762	1.556
Cys	1.512	-0.233	0.592	0.488	0.130	0.161	0.327	1.088	0.425	0.239	0.346	1.140	-0.162	0.980	-0.425	0.368
Asp	1.637	-0.332	2.772	-1.711	1.225	1.257	2.051	-0.708	-0.061	0.753	0.088	1.664	-0.917	1.640	-1.126	-0.313
Glu	0.351	0.880	-0.467	1.574	0.035	0.066	1.057	0.436	-0.270	0.954	0.872	0.561	0.660	0.230	1.033	1.810
Phe	1.267	0.035	1.384	-0.321	-0.449	-0.417	0.235	1.137	0.597	0.051	-0.022	1.573	-0.004	0.771	0.518	1.324
Gly	1.244		1.100		-0.031		1.444		0.763		1.367		0.880		-0.780	
His	0.367	0.999	1.050	-0.029	-0.469	-0.437	0.261	0.978	1.249	-0.410	1.159	0.496	-0.001	0.709	0.709	1.533
Ile	-0.217	1.463	0.261	0.834	0.058	0.090	1.125	0.327	1.263	-0.589	0.560	0.749	0.463	0.406	0.313	1.094
Lys	1.441		0.979		-0.032		1.101		0.999		1.934		0.597		-0.854	
Leu	0.997	0.285	1.284	-0.209	-0.824	-0.792	1.577	-0.192	-0.184	0.948	1.055	0.575	-0.048	0.847	-0.979	-0.182
Met	1.497	-0.195	1.556	-0.493	3.228	3.260	1.341	-0.027	0.054	0.985	3.357	-1.724	-2.075	2.932	-0.047	0.735
Asn	0.609	0.630	2.218	-1.120	1.509	1.540	0.397	1.068	0.621	0.011	0.316	1.040	1.550	-0.693	0.038	0.825
Pro	-0.015	1.340	0.316	0.732	-0.189	-0.158	1.268	0.068	-0.669	1.365	0.169	1.648	0.456	0.249	0.403	1.225
Gln	1.607	-0.328	1.409	-0.332	-0.659	-0.628	-0.450	1.869	0.333	0.296	0.106	1.369	0.221	0.594	0.339	1.133
Arg	0.862	0.465	0.966	0.082	0.335	0.367	1.154	0.146	1.069	-0.229	1.396	0.186	0.214	0.549	0.136	0.946
Ser	2.636	-1.368	0.624	0.466	0.257	0.288	2.142	-0.733	1.393	-0.658	0.837	0.554	0.167	0.669	0.469	1.258
Thr	0.576	0.771	1.777	-0.747	-0.112	-0.080	-0.388	1.638	-0.060	1.096	3.200	-1.489	-0.129	0.920	-0.220	0.584
Val	0.022	1.327	0.704	0.327	0.038	0.070	0.451	0.826	1.382	-0.647	0.318	1.292	-0.468	1.179	0.454	1.274
Trp	1.538	-0.154	0.928	0.086	-1.325	-1.293	0.982	0.208	3.371	-2.464	1.523	-0.087	2.649	-1.953	-0.319	0.516
Tyr		1.264		1.085		0.031		1.451		0.606		1.483		0.830		0.789
(Const)	-54.961	-24.160	-65.818	0.800	-5.439	-7.370	-61.689	-21.392	-34.662	-11.499	-56.009	-38.837	-12.554	-35.654	-0.636	-49.515

3.3.2. Aminoácidos asociados mediante el uso de codones con las clasificaciones taxonómicas en archaeabacterias, bacterias y eucariotes.

Usando el método CHAID en el caso de los tres reinos pero para el caso donde sean secuencias de uso de genes en una base de datos que muestra diversidad de organismos y tipos de proteinas presentes, se obtienen resultados excelentes desde el punto de vista de clasificación así como interacción entre aminoácidos lo cual se muestra en la Tabla 3.3.2.1, donde la Serina alcanza un 98,7 % de clasificación, Tabla 3.3.2.1, y siendo el que mayor significación posee, apareciendo en el nodo principal del árbol de la Figura3.3.2.1, donde aparecen en los nodos secundarios aminoácidos como la Lisina y el Triptófano que también presentan una buena significación.

Tabla 3.3.2.1. Significación de los aminoácidos al ser utilizados como variables predictoras en la construcción de árboles de decisión y los porcientos de clasificación alcanzados.

AA	Sig.	%clasificación
Leucina	1.03509E-09	98
Ácido Aspártico	4.12693E-11	98
Metionina	1.77809E-11	96,9
Fenilalanina	6.98123E-13	97,3
Triptófano	1.86E-14	98,2
Arginina	7.5487E-18	96,7
Tirosina	4.05541E-20	97,1
Prolina	1.48704E-28	98,4
Asparagina	7.28407E-29	98,4
Treonina	5.98501E-39	97,6
Ácido Glutámico	1.77E-40	97,3
Histidina	1.61649E-46	97,6
Glicina	8.24765E-48	96,2
Lisina	2.03555E-51	98,2
Isoleucina	3.92231E-57	98,9
Glutamina	9.01E-59	98,4
Valina	6.9077E-64	97,8
Cisteína	5.6843E-64	97,1
Alanina	7.97147E-66	97,1
Serina	1.9151E-99	98,7

[a]Sig. Significación del estadígrafo de razón verosimilitud Chi-cuadrado. Por simplificación se ha utilizado el simbolismo del SPSS para la notación científica, es decir, por ejemplo, el símbolo E-05 significa 10^{-5}.

Tabla 3.3.2.2. Clasificación obtenida con método CHAID en la bases de datos curada con validación cruzada.

Muestras Observadas	Predicciones			
	archaeas	baterias	eucariotes	Exactitud
archaeas	45	2	3	90.0%
baterias	0	49	1	98.0%
eucariotes	0	0	350	100.0%
% Total	10.0%	11.3%	78.7%	98.7%

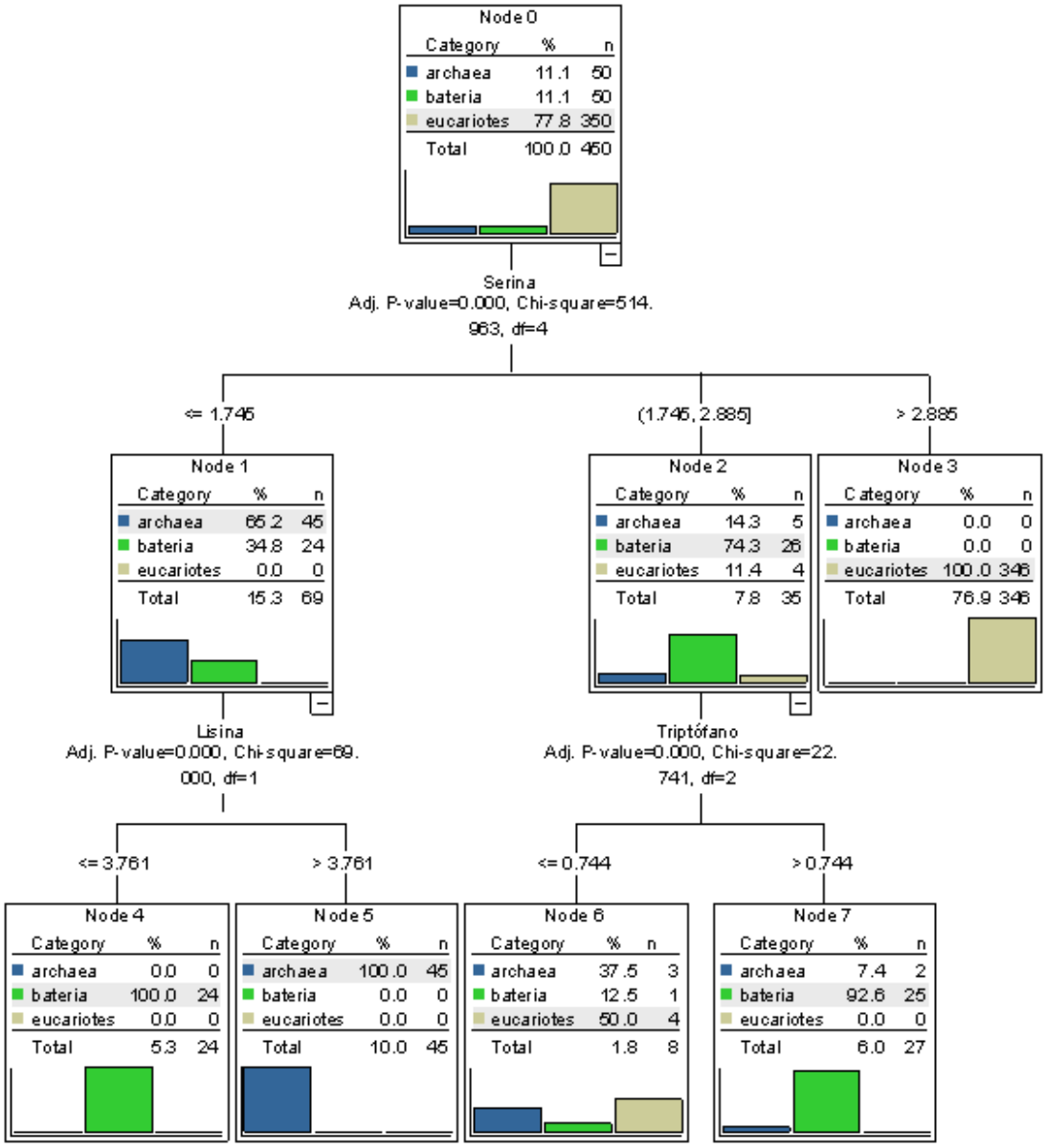


Figura 3.3.2.1 Árbol de aminoácidos asociados con los resultados en la base de datos curada con validación cruzada en las clasificaciones taxonómicas de archaeas, bacterias y eucariote.

Tabla 3.3.2.3. Clasificación obtenida con método CHAID en la base de datos tomando aleatoriamente el 70% de la base como entrenamiento y el resto usado en validación externa.

		Predicciones			
muestra	observada	archaeas	bacterias	eucariotes	Exactitud
entrenamiento	archaeas	38	1	0	97.4%
	bacterias	1	36	0	97.3%
	eucariotes	0	0	33	100.0%
	% Total	35.8%	33.9%	30.3%	98.2%
validac	archaeas	10	1	0	90.9%
	bacterias	0	13	0	100.0%
	eucariotes	0	0	17	100.0%
	% Total	24.4%	34.1%	41.5%	97.6%

3.3.2.1. Análisis de Discriminante y la evaluación del desempeño de los clasificadores.

Con el análisis de discriminante realizado en esta taxa se comprueba que todos los aminoácidos están asociados con la clasificación de los vectores NEC_k en los tres reinos. En la Tabla 3.3.2.1.1 se puede ver que, el aminoácido Tirosina está incluído solo en las combinaciones lineales de las funciones discriminantes cuando se utiliza el método Stepwise. Mientras, en la Tabla 3.3.2.1.2 se puede apreciar que la eficacia de las funciones discriminantes en la separación de los casos en grupos, expresada a través de las correlaciones canónicas, es similar para ambos procedimientos. Además, los valores de la Lambda de Wilk y la significación del test Chi-cuadrado indican que las capacidades discriminatorias de las funciones obtenidas por estos procedimientos son similares. En particular, para todas las funciones los valores de estos parámetros son altos, indicando el buen desempeño de las funciones discriminantes, que también se observa en el gráfico de dispersión que aparece en la Figura 3.3.2.1.1.

En la comparación de los clasificadores no hay diferencias en los indicadores este hecho se ilustra en las curvas ROC obtenidas (Figura 3.3.2.1.2) y en la Tabla 3.3.2.1.3, en la que se muestra que los intervalos de confianza asimtóticos para 95% de confianza de las áreas bajo la curva ROC. Cuando se utilizan los parámetros derivados de la matríz de confusión para evaluar el desempeño de estos clasificadores, nos sugieren que las diferencias entre

los clasificadores no son significativas. En la Tabla 3.3.2.1.4 se muestran los valores de los parámetros mencionados.

Tabla 3.3.2.1.1. Funciones discriminantes canónicas obtenidas con la introducción de todos los aminoácidos que satisfacen el test de tolerancia y con el método Stepwise.

Aminoácido	Todas		Stepwise	
	Función discriminante		Función discriminante	
	1	2	1	2
Alanina	1.87522537	2.71219483	-	-
Cisteína	2.35349121	1.18627654	-	-
Ácido Aspártico	1.65542172	3.79086839	0.01505919	2.05921876
Ácido Glutámico	0.87853914	1.21141879	0.76471639	1.39716935
Fenilalanina	1.33390973	2.39542641	-	-
Glicina	2.0104053	2.14675602	-	-
Histidina	3.00759821	4.41305899	-	-
Isoleucina	0.64295095	2.41513135	1.17796651	0.27728312
Lisina	1.32325793	1.7472425	0.53833916	0.71379549
Leucina	1.18522417	2.86236885	0.64675298	0.44148571
Metionina	1.68716796	2.00676784	-	-
Asparagina	1.301334	3.1603493	-	-
Prolina	0.79049656	1.32850812	2.25283399	1.36749052
Glutamina	2.81106128	3.52604836	1.10296299	1.15205197
Arginina	0.66028745	1.79338244	0.82580626	0.59797662
Serina	4.30810311	0.83123861	2.53607525	1.61273661
Treonina	0.942135	3.88635571	1.02379164	1.61152628
Valina	0.13613422	2.18525549	1.48574882	0.39439735
Tirosina	-	-	2.25373227	2.07990895
Triptófano	0.34218149	2.18638374	-	-
(Constante)	78.0678216	141.079972	29.0273106	9.24919259

Tabla 3.3.2.1.2. Eficacia de las funciones discriminantes a través de las correlaciones canónicas y los valores de la Lambda de Wilk.

Función	Valor principal	% de Varianza	% Var. Acum.	Corr. Canónica	Función	Lambda de Wilks	Chi cuadrado	g.l.	Sig.
Stepwise									
1	36.883	73.164	73.164	0.987	1 a 2	0.002	634.217	24	0.000
2	13.529	26.836	100.000	0.965	2	0.069	268.950	11	0.000
Todas las variables									
1	39.824	71.749	71.749	0.988	1 a 2	0.001	632.783	38	0.000
2	15.681	28.251	100.000	0.970	2	0.060	272.983	18	0.000

Tabla 3.3.2.1.3. Resultado del área bajo la curva en los tres métodos utilizados.

Resultados del Análisis	Área	Error Estándar	Sig. Asintótica	Intervalo de confianza asintótico para el 95%	
				Límite superior	Límite inferior
Probabilidad Archaea (Análisis Disc. Stepwise)	1.000	0.000	0.000	1.000	1.000
Probabilidad Archaea (Análisis CHAID)	0.997	0.002	0.000	0.993	1.002
Probabilidad Archaea (Análisis Disc. Todas)	1.000	0.000	0.000	1.000	1.000
Probabilidad Bacteria (Análisis Disc. Stepwise)	1.000	0.000	0.000	1.000	1.000
Probabilidad Bacteria (Análisis CHAID)	0.997	0.002	0.000	0.993	1.002
Probabilidad Bacteria (Análisis Disc. Todas)	1.000	0.000	0.000	1.000	1.000
Probabilidad Eucariotes (Análisis Disc. Stepwise)	1.000	0.000	0.000	1.000	1.000
Probabilidad Eucariotes (Análisis CHAID)	1.000	0.000	0.000	1.000	1.000
Probabilidad Eucariotes (Análisis Disc. Todas)	1.000	0.000	0.000	1.000	1.000

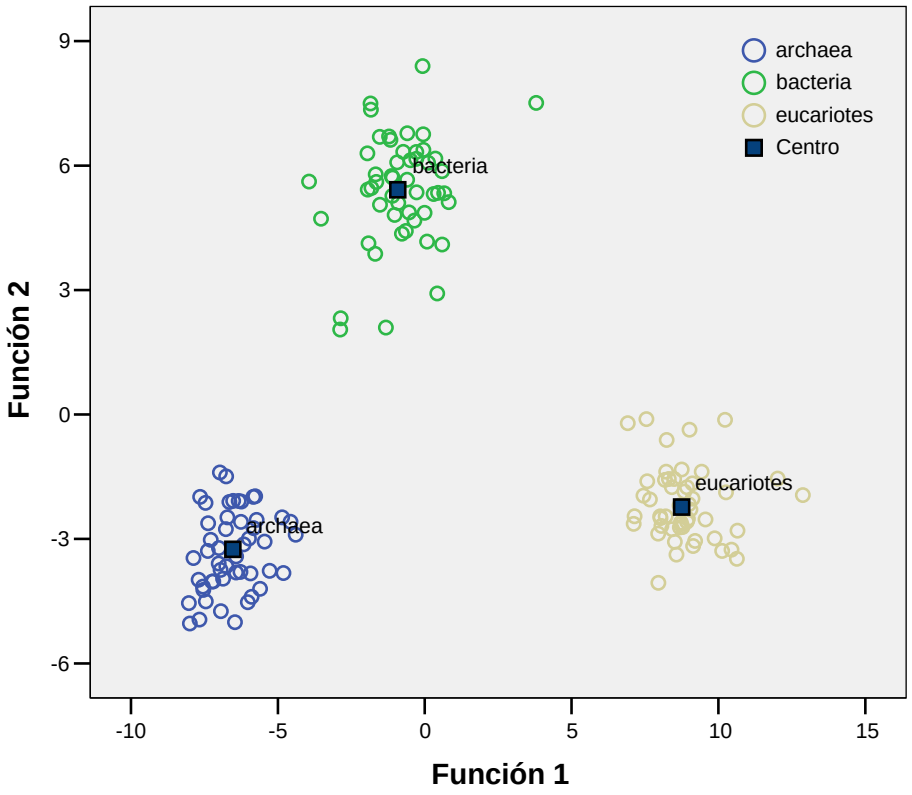


Figura 3.3.2.1.1 Gráfico de dispersión de la función Discriminante.

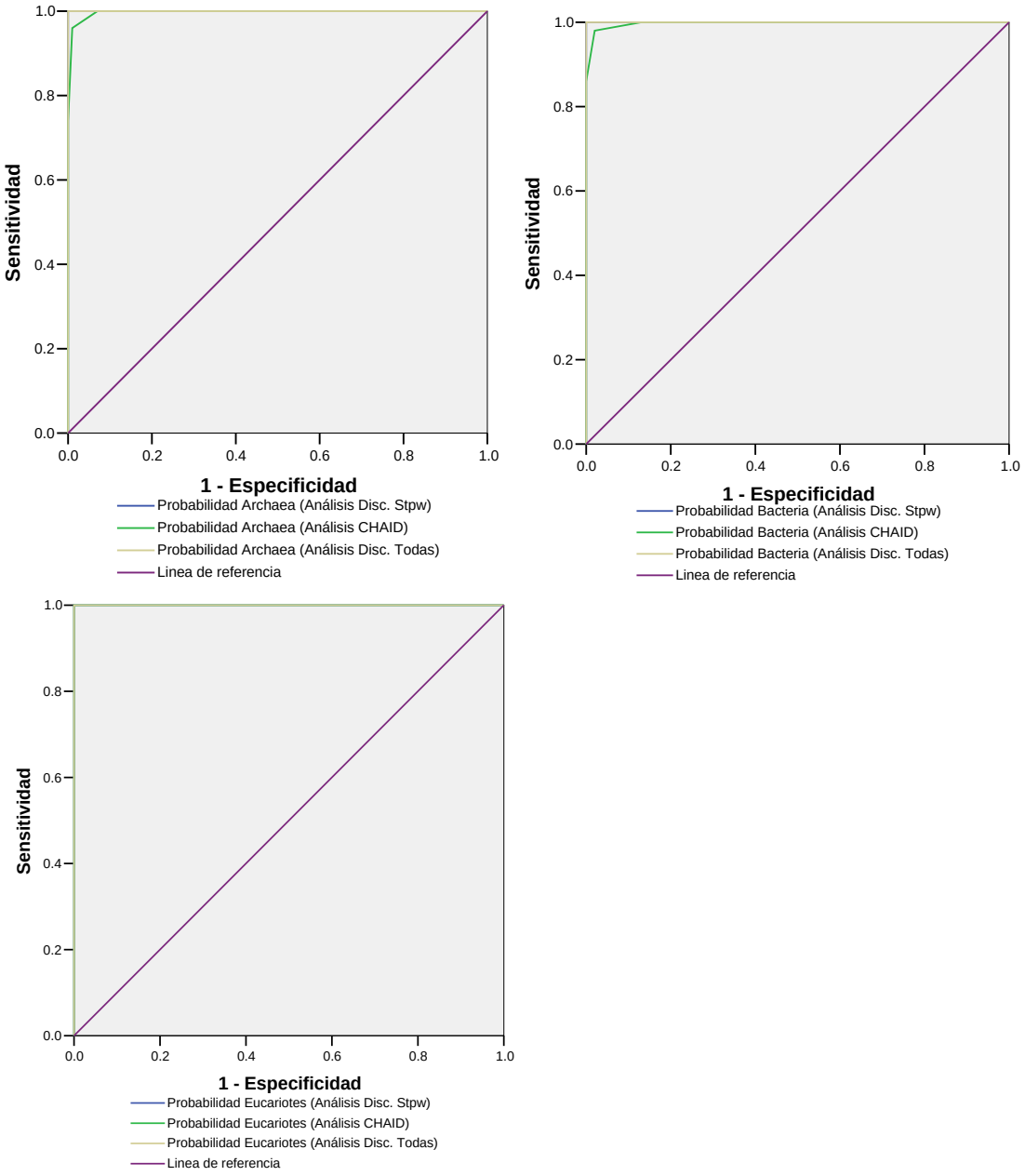


Figura 3.3.2.1.2 Curvas ROC obtenidas con los dos métodos de discriminante y con el método CHAID.

Tabla 3.3.2.1.4 Parámetros calculados a partir de la matriz de confusión para evaluar el desempeño de los clasificadores utilizados.

Predicciones de los miembros del Grupo con Anl. Discriminante Stepwise					
70 % base de datos extendida					
Org.	**Razón de TP**	**Razón de TN**	**Precisión**	**Exactitud**	**% Clasif.**
Archaea	100.0	100.0	100.0	100.0	100.0
Bacteria	100.0	100.0	100.0	100.0	100.0
Eucariotes	100.0	100.0	100.0	100.0	100.0
Validación cruzada					
Archaea	100.0	98.6	97.5	99.1	100.0
Bacteria	97.3	100.0	100.0	99.1	97.3
Eucariotes	100.0	98.7	100.0	99.1	100.0
Validación externa					
Archaea	100.0	100.0	100.0	100.0	100.0
Bacteria	100.0	100.0	100.0	100.0	100.0
Eucariotes	100.0	100.0	100.0	100.0	100.0
Predicciones de los miembros del Grupo con Anl. Discriminante (todas)					
70 % base de datos extendida					
Archaea	100.0	100.0	100.0	100.0	100.0
Bacteria	100.0	100.0	100.0	100.0	100.0
Eucariotes	100.0	100.0	100.0	100.0	100.0
Validación cruzada					
Archaea	100.0	98.6	97.5	99.1	100.0
Bacteria	97.3	100.0	100.0	99.1	97.3
Eucariotes	100.0	98.7	100.0	99.1	100.0
Validación externa					
Archaea	100.0	100.0	100.0	100.0	100.0
Bacteria	100.0	100.0	100.0	100.0	100.0
Eucariotes	100.0	100.0	100.0	100.0	100.0
Predicciones de los miembros del Grupo con CHAID					
70 % base de datos extendida					
Archaea	97.4	98.6	97.4	98.2	97.4
Bacteria	97.3	98.6	97.3	98.2	97.3
Eucariotes	100.0	97.4	100.0	98.2	100.0
Validación externa					
Archaea	90.9	100.0	100.0	97.6	90.9
Bacteria	100.0	96.4	92.9	97.6	100.0
Eucariotes	100.0	95.8	100.0	97.6	100.0

3.3.3. Aminoácidos asociados mediante el uso de codones con las clasificaciones taxonómicas en archaeabacterias y bacterias.

Con el método CHAID en estos dos reinos, se observa en la Tabla 3.3.3.1, que al igual que en la taxa anterior el aminoácido Serina tiene el mejor porciento de clasificación, mientras la mayor significación la posee la Lisina que también estaba presente en los aminoácidos de mayor significación en la taxa anterior, los porcientos de clasificación se pueden ver en la Tabla 3.3.3.2 y el árbol correspondiente es el que aparece en la Figura 3.3.3.1.

Tabla 3.3.3.1. Significación de los aminoácidos al ser utilizados como variables predictoras en la construcción de árboles de decisión y los porcientos de clasificación alcanzados.

AA	Sig.	%clasificación
Asparagina	0.01478413	96
Cisteína	4.81751E-05	98
Leucina	1.4552E-05	96
Valina	1.26618E-06	95
Ácido Aspártico	1.36782E-07	94
Metionina	1.04092E-07	96
Arginina	2.40979E-08	97
Serina	1.04247E-08	98
Triptófano	9.33E-13	94
Tirosina	2.14801E-15	96
Glicina	1.86287E-15	96
Prolina	7.69908E-21	97
Histidina	1.09352E-21	94
Treonina	1.33227E-22	97
Ácido Glutámico	5.42E-24	96
Fenilalanina	5.41845E-24	95
Isoleucina	5.41845E-24	96
Alanina	1.28041E-24	96
Glutamina	4.21E-25	97
Lisina	1.33286E-26	96

[a]Sig. Significación del estadígrafo de razón verosimilitud Chi-cuadrado. Por simplificación se ha utilizado el simbolismo del SPSS para la notación científica, es decir, por ejemplo, el símbolo E-05 significa 10^{-5}.

Tabla 3.3.3.2. Clasificación obtenida con método CHAID en la bases de datos curada con validación cruzada.

Muestras Observadas	Predicciones		
	archaeas	bacterias	Exactitud
archaeas	50	0	100.0%
bacterias	4	46	92.0%
% Total	54.0%	46.0%	96.0%

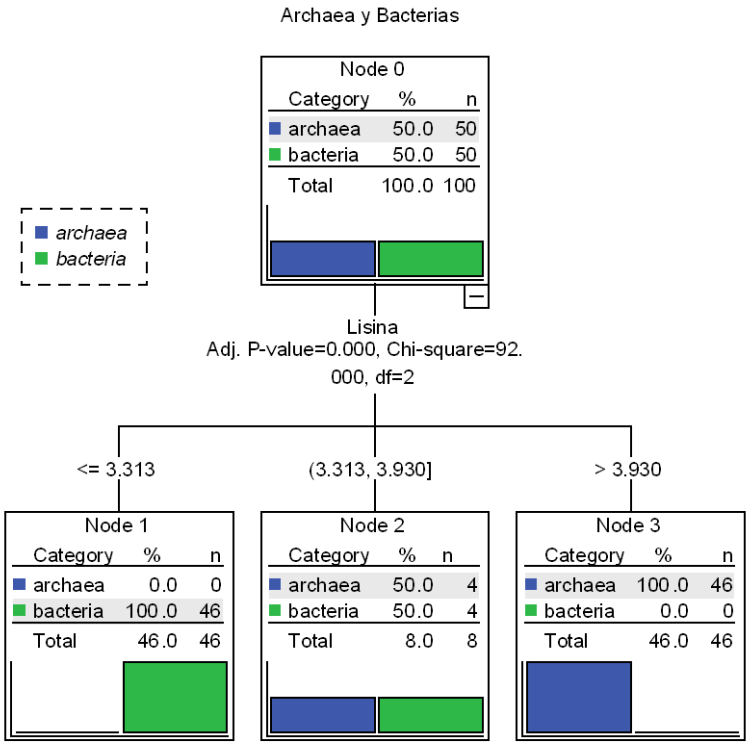


Figura 3.3.3.1 Árbol de Aminoácidos asociados con los resultados en la base de datos curada con validación cruzada en las clasificaciones taxonómicas de archaeas y bacterias.

Tabla 3.3.3.3. Clasificación obtenida con método CHAID en la base de datos tomando aleatoriamente el 70% de la base como entrenamiento y el resto usado en validación externa.

		Prediccionesd		
muestra	Observada	archaea	bacterias	Exactitud
Entrena miento	archaea	39	0	100.0%
	bacterias	2	35	94.6%
	% Total	53.9%	46.1%	97.4%
Validac	archaea	11	0	100.0%
	bacterias	1	12	92.3%
	% Total	50.0%	50.0%	95.8%

3.3.3.1. Análisis de Discriminante y la evaluación del desempeño de los clasificadores.

Con el análisis de discriminante realizado en esta taxa en la Tabla 3.3.3.1.1 se puede ver que, el aminoácido tirosina es el único que no aparece en el método cuando entran todas las variables, que superan el test de tolerancia, mientras que cuando se ejecuta el método con la variante Stepwise solo intervienen seis aminoácidos.

Mientras, en la Tabla 3.3.3.1.2 se puede apreciar que la eficacia de las funciones discriminantes en la separación de los casos en grupos, expresada a través de las correlaciones canónicas, es similar para ambos procedimientos. Además, los valores de la Lambda de Wilk y la significación del test Chi-cuadrado indican que las capacidades discriminatorias de las funciones obtenidas por estos procedimientos son similares.

Los resultados de clasificación global no son estadísticamente diferentes para los métodos de obtención de las funciones discriminantes y para el método CHAID. Este hecho se ilustra en las curvas ROC obtenidas (Figura 3.3.3.1.1) y en la Tabla 3.3.3.1.3, en la que se muestra los indicadores de las áreas bajo la curva ROC. Al utilizar los parámetros derivados de la matríz de confusión para evaluar el desempeño de estos clasificadores, observamos que las diferencias no son significativas. En la Tabla 3.3.3.1.4 se muestran los valores de los parámetros mencionados.

Tabla 3.3.3.1.1. Funciones discriminantes canónicas obtenidas con la introducción de todos los aminoácidos que satisfacen el test de tolerancia y con el método Stepwise.

Aminoácidos	Función discriminante	
	Todas	Stepwise
Alanina	4.62486353	-
Cisteína	3.40866802	-
Ácido Aspártico	4.35329008	-
Ácido Glutámico	3.01312637	0.87285347
Fenilalanina	4.69645842	-
Glicina	4.75110106	-
Histidina	8.76823771	2.83974481
Isoleucina	4.40842095	-
Lisina	3.10784031	0.95375833
Leucina	3.62521345	-
Metionina	3.67463302	-
Asparagina	6.09094983	2.73620833
Prolina	2.35739495	-
Glutamina	6.27960072	2.3650078
Arginina	3.79944512	-
Serina	4.48932266	-
Treonina	4.5272397	-
Valina	3.02606723	-
Tirosina	-	3.71432475
Triptófano	2.84141122	-
(Constant)	242.335636	1.59086365

Tabla 3.3.3.1.2. Eficacia de las funciones discriminantes a través de las correlaciones canónicas y los valores de la Lambda de Wilk.

Función	Valor principal	% de Varianza	% Var. Acum.	Corr. Canónica	Función	Lambda de Wilks	Chi cuadrado	g.l.	Sig.
Stepwise									
1	25.143	100	100	0.981	1	0.038	231.714	6	0.000
Todas las variables									
1	32.793	100	100	0.985	1	0.030	227.057	19	0.000

Tabla 3.3.3.1.3. Resultado del área bajo la curva en los tres métodos utilizados.

Resultados del Análisis	Área	Error Estándar	Sig. Asintótica	Intervalo de confianza asintótico para el 95%	
				Límite superior	Límite inferior
Probabilidad Bacteria (Análisis Disc. Stepwise)	1.000	0.000	0.000	1.000	1.000
Probabilidad Bacteria (Análisis CHAID)	0.970	0.020	0.000	0.931	1.009
Probabilidad Bacteria (Análisis Discriminante)	1.000	0.000	0.000	1.000	1.000

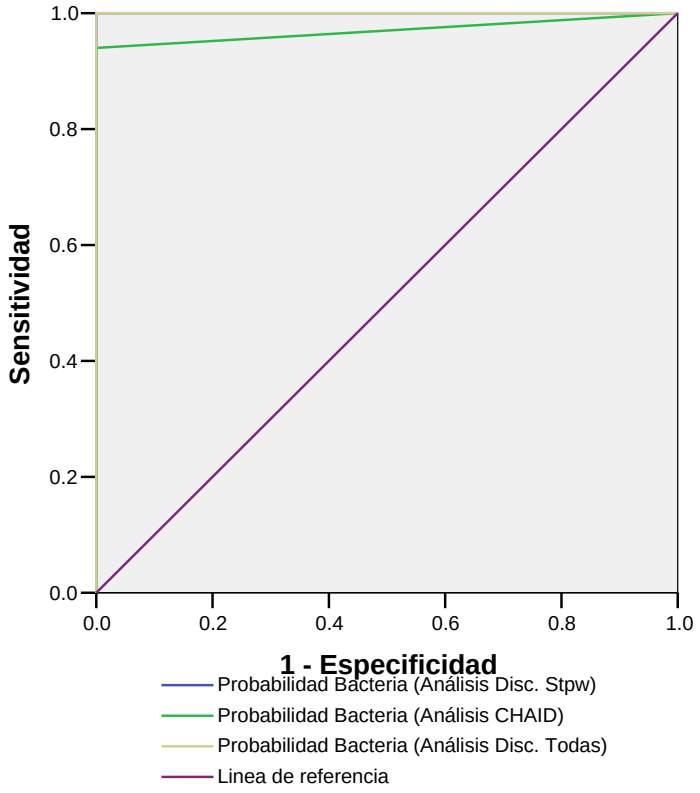


Figura 3.3.3.1.1 Curvas ROC obtenidas con los dos métodos de discriminante y con el método CHAID.

Tabla 3.3.3.1.4 Parámetros calculados a partir de la matriz de confusión para evaluar el desempeño de los clasificadores utilizados.

	Grupos	Razón de TP	Razón de TN	Prec.	Exac.	% de Clasf.
Predicciones de los miembros del Grupo con Anl. Discriminante (Stpw).						
70 % base de datos extendida	**Archaea**	100.0	100.0	100.0	100.0	100.0
	Bacteria	100.0	100.0	100.0	100.0	100.0
Validación cruzada	**Archaea**	100.0	100.0	100.0	100.0	100.0
	Bacteria	100.0	100.0	100.0	100.0	100.0
Validación externa	**Archaea**	100.0	100.0	100.0	100.0	100.0
	Bacteria	100.0	100.0	100.0	100.0	100.0
Predicciones de los miembros del Grupo con Anl. Discriminante (Todas).						
70 % base de datos extendida	**Archaea**	100.0	100.0	100.0	100.0	100.0
	Bacteria	100.0	100.0	100.0	100.0	100.0
Validación cruzada	**Archaea**	100.0	97.3	97.5	98.7	100.0
	Bacteria	97.3	100.0	100.0	98.7	97.3
Validación externa	**Archaea**	100.0	100.0	100.0	100.0	100.0
	Bacteria	100.0	100.0	100.0	100.0	100.0
Predicciones de los miembros del Grupo con CHAID						
70 % base de datos extendida	**Archaea**	100.0	94.6	95.1	97.4	100.0
	Bacteria	94.6	100.0	100.0	97.4	94.6
Validación externa	**Archaea**	100.0	92.3	91.7	95.8	100.0
	Bacteria	92.3	100.0	100.0	95.8	92.3

3.3.4. Aminoácidos asociados mediante el uso de codones con las clasificaciones taxonómicas en vertebrados e invertebrados.

El análisis realizado en esta taxa nos proporcionó los datos que aparecen la Tabla 3.3.4.1, donde podemos ver que el aminoácido Asparagina que alcanza mayor porciento de clasificación con validación cruzada. Mientras la Leucina es la de mayor significación aplicando este mismo método.

Tabla 3.3.4.1. Significación de los aminoácidos al ser utilizados como variables predictoras en la construcción de árboles de decisión y los porcientos de clasificación alcanzados.

AA	Sig.	%clasificación
Glicina	0.048974524	91,3
Ácido Glutámico	0.00427	92
Valina	0.001548906	89
Metionina	0.00087178	91,7
Fenilalanina	6.84314E-05	92,7
Lisina	6.33727E-05	91,7
Prolina	1.14097E-05	90,7
Arginina	9.60722E-06	90,7
Alanina	6.9889E-06	90,7
Histidina	4.03007E-06	90,7
Isoleucina	2.63794E-06	93
Treonina	9.61E-07	90,7
Cisteína	2.88813E-08	91
Glutamina	1.47E-09	94
Tirosina	9.34572E-10	92,3
Serina	1.80376E-10	93
Triptófano	7.40E-11	90,3
Ácido Aspártico	2.5637E-21	92,7
Asparagina	2.12837E-31	94,7
Leucina	5.6054E-32	92,3

[a]Sig. Significación del estadígrafo de razón verosimilitud Chi-cuadrado. Por simplificación se ha utilizado el simbolismo del SPSS para la notación científica, es decir, por ejemplo, el símbolo E-05 significa 10^{-5}.

Tabla 3.3.4.2. Clasificación obtenida con método CHAID en la bases de datos curada con validación cruzada.

Muestra Observada	Predicciones		
	vertebrados	invertebrados	Exactitud
vertebrados	184	16	92.0%
invertebrados	7	93	93.0%
% Total	63.7%	36.3%	92.3%

Tabla 3.3.4.3. Clasificación obtenida con método CHAID en la base de datos tomando aleatoriamente el 70% de la base como entrenamiento y el resto usado en validación externa.

		Predicciones		
muestra	observada	vert	invert	Exactitud
entrena miento	vertebrados	65	7	90.3%
	invertebrados	5	71	93.4%
	% Total	47.3%	52.7%	91.9%
validac	vertebrados	21	7	75.0%
	invertebrados	4	20	83.3%
	% Total	48.1%	51.9%	78.8%

3.3.4.1. Análisis de Discriminante y la evaluación del desempeño de los clasificadores.

Con el análisis de discriminante realizado se obtinen los resultados que aparecen en la Tabla 3.3.4.1.1 donde se presentan las funciones discriminantes obtenidas por el método Stepwise minimizando la Lambda de Wilk y sin aplicar este método considerando que entren todas las que superen el test de tolerancia, en este caso como podemos observar solo una la tirosina no entra, mientra en el método de Stepwise solo participan 7 aminoácidos.

Mientras, en la Tabla 3.3.4.1.2 se puede apreciar que la eficacia de las funciones discriminantes en la separación de los casos en grupos, expresada a través de las correlaciones canónicas, es similar para ambos procedimientos. Además, los valores de la Lambda de Wilk y la significación del test Chi-cuadrado indican que las capacidades discriminatorias de las funciones obtenidas por estos procedimientos son similares.

El hecho de que no haya diferencias estadísticamente detectables en los metodos de Discriminante y CHAID se ilustra en las curvas ROC obtenidas (Figura 3.3.4.1.1) y en la Tabla 3.3.4.1.3, donde aparecen los parámetros que describen las áreas bajo la curva ROC. Sin embargo, al utilizar los parámetros derivados de la matríz de confusión para evaluar el desempeño de estos clasificadores, nos sugieren que existen algunas diferencias entre los clasificadores. En la Tabla 3.3.4.1.4 se muestran los valores de los parámetros mencionados.

Tabla 3.3.4.1.1. Funciones discriminantes canónicas obtenidas con la introducción de todos los aminoácidos que satisfacen el test de tolerancia y con el método Stepwise.

Aminoácidos	Función discriminante	
	Todas	Stepwise
Alanina	-1.153408	1.3635172
Cisteína	0.6291342	-
Ácido Aspártico	-1.283027	1.3734601
Ácido Glutámico	1.3230351	-0.501296
Fenilalanina	0.4843401	-
Glicina	-0.091928	-
Histidina	0.8533897	-
Isoleucina	0.1842257	-
Lisina	0.11743	-
Leucina	1.0027172	-0.790415
Metionina	-2.053798	2.2627063
Asparagina	-1.582589	1.9864618
Prolina	0.7630663	-
Glutamina	-0.440902	-
Arginina	0.1671537	-
Serina	0.69771	-0.447447
Treonina	-0.036089	-
Valina	0.6549796	-
Tirosina	-	-
Triptófano	0.1301897	-
(Constant)	-6.046161	-10.44801

Tabla 3.3.4.1.2. Eficacia de las funciones discriminantes a través de las correlaciones canónicas y los valores de la Lambda de Wilk.

Función	Valor principal	% de Varianza	% Var. Acum.	Corr. Canónica	Función	Lambda de Wilks	Chi cuadrado	g.l.	Sig.
Stepwise									
1	3.364	100	100	0.878	1	0.229	209.965	7	0.000
Todas las variables									
1	3.841	100	100	0.891	1	0.207	215.275	19	0.000

Tabla 3.3.4.1.3. Resultado del área bajo la curva en los tres métodos utilizados.

Resultados del Análisis	Área	Error Estándar	Sig. Asintótica	Intervalo de confianza asintótico para el 95%	
				Límite superior	Límite inferior
Probabilidad Vertebrados (Análisis Disc. Stepwise)	0.996	0.002	0.000	0.991	1.000
Probabilidad Vertebrados (Análisis CHAID)	0.945	0.017	0.000	0.912	0.978
Probabilidad Vertebrados (Análisis Disc. Todas)	0.990	0.007	0.000	0.977	1.003

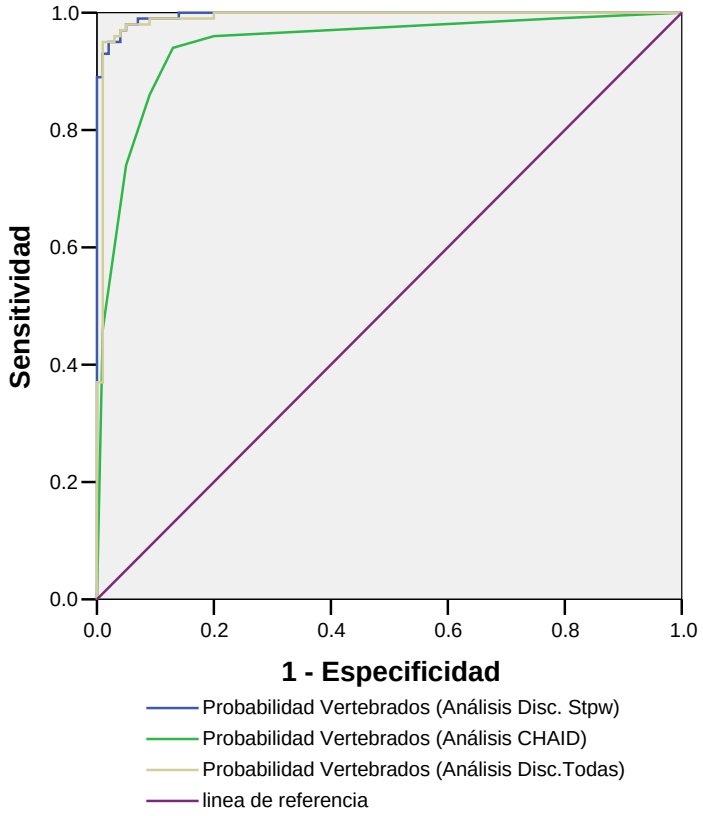


Figura 3.3.4.1.1 Curvas ROC obtenidas con los dos métodos de discriminante y con el método CHAID.

Tabla 3.3.4.1.4 Parámetros calculados a partir de la matriz de confusión para evaluar el desempeño de los clasificadores utilizados.

	Grupos	Razón de TP	Razón de TN	Prec.	Exac.	% de Clasf.
Predicciones de los miembros del Grupo con Anl. Discriminante (Stpw).						
70 % base de datos extendida	**Vert**	94.4	97.4	97.1	95.9	94.4
	Invert	97.4	94.4	94.9	95.9	97.4
Validación cruzada	**Vert**	93.1	96.1	95.7	94.6	93.1
	Invert	96.1	93.1	93.6	94.6	96.1
Validación externa	**Vert**	96.4	95.8	96.4	96.2	96.4
	Invert	95.8	96.4	95.8	96.2	95.8
Predicciones de los miembros del Grupo con Anl. Discriminante (Todas).						
70 % base de datos extendida	**Vert**	95.8	97.4	97.2	96.6	95.8
	Invert	97.4	95.8	96.1	96.6	97.4
Validación cruzada	**Vert**	95.8	96.1	95.8	95.9	95.8
	Invert	96.1	95.8	96.1	95.9	96.1
Validación externa	**Vert**	92.9	95.8	96.3	94.2	92.9
	Invert	95.8	92.9	92.0	94.2	95.8
Predicciones de los miembros del Grupo con CHAID						
70 % base de datos extendida	**Vert**	90.3	93.4	92.9	91.9	90.3
	Invert	93.4	90.3	91.0	91.9	93.4
Validación externa	**Vert**	75.0	83.3	84.0	78.8	75.0
	Invert	83.3	75.0	74.1	78.8	83.3

3.3.5. Aminoácidos asociados mediante el uso de codones con las clasificaciones taxonómicas en vertebrados no mamíferos y mamíferos.

Cuando se aplica la técnica CHAID a vectores que expresan probabilidad de frecuencia en el uso de codones en estos dos grupos de organismos tan cercanos en los aspectos que los caracterizan desde el punto de vista evolutivo, los resultados obtenidos de la base de datos curada con una validación cruzada, Tabla 3.3.5.1, muestran que al igual que para las secuencias de aminoácidos la Metionina es la que mejor significación tiene, mientras que la Leucina es la de mayor porciento de clasificación. Podemos señalar que en este caso dos aminoácidos no alcanzan valores menores que 0.05 en su significación ellos son la

Fenilalanina y el Ácido Glutámico. En la tabla 3.3.5.2, se observan los porcientos que se obtienen al realizar una validación del 70% de la base curada como entrenamiento con una validación externa con el resto de la base.

Tabla 3.3.5.1. Significación de los aminoácidos al ser utilizados como variables predictoras en la construcción de árboles de decisión y los porcientos de clasificación alcanzados.

AA	Sig.[a]	%clasificación
Glutamina	3.2514E-05	88,5
Valina	6.6791E-06	88,5
Cisteína	0.000192	87
Treonina	6.7907E-06	90
Tirosina	0.002	87,5
Prolina	0.004356	87
Histidina	0.000184	89
Isoleucina	0.031889439	88,5
Arginina	0.003529707	90
Lisina	0.001127208	91,5
Glicina	0.000644454	89
Ácido Aspártico	2.9716E-05	91
Triptófano	3.39E-06	89
Leucina	1.36446E-08	93,5
Asparagina	1.33436E-08	91
Serina	4.77861E-09	90
Alanina	1.76108E-10	90
Metionina	6.9716E-11	90,5

[a]Sig. Significación del estadígrafo de razón verosimilitud Chi-cuadrado. Por simplificación se ha utilizado el simbolismo del SPSS para la notación científica, es decir, por ejemplo, el símbolo E-05 significa 10^{-5}.

Tabla 3.3.5.2. Clasificación obtenida con método CHAID en la base de datos tomando aleatoriamente el 70% de la base como entrenamiento y el resto usado en validación externa.

		Predicciones		
muestra	Observada	vertNoMamif	Mamiferos	Percent Correct
entrenamiento	vertNoMamif	66	6	91.7%
	Mamiferos	5	69	93.2%
	% Total	48.6%	51.4%	92.5%
validac	vertNoMamif	22	6	78.6%
	Mamiferos	4	22	84.6%
	% Total	48.1%	51.9%	81.5%

3.3.5.1. Análisis de Discriminante y la evaluación del desempeño de los clasificadores.

El análisis de discriminante realizado en esta taxa, Tabla 3.3.5.1.1, donde se describen las funciones discriminantes canónicas obtenidas con la introducción de todos los aminoácidos que satisfacen el test de tolerancia y con el método Stepwise, para el cual solo intervienen 8 aminoácidos. Se pude señalar en el caso del aminoácido Tirosina no aparece en ninguno de los dos métodos aplicados.

En la Tabla 3.3.5.1.2 se puede apreciar que la eficacia de las funciones discriminantes en la separación de los casos en grupos, expresada a través de las correlaciones canónicas, es similar para ambos procedimientos. Además, los valores de la Lambda de Wilk y la significación del test Chi-cuadrado indican que las capacidades discriminatorias de las funciones obtenidas por estos procedimientos son similares.

En las curvas ROC obtenidas (Figura 3.3.5.1.1), que el análisis Discriminante realizado es superior en sus dos variantes al CHAID y en la Tabla 3.3.5.1.3, en la que se muestra que los intervalos de confianza asimtóticos para 95% de confianza de las áreas bajo la curva ROC también se observan que los mejores indicadores se refieren a la técnica Discriminante. Al utilizar los parámetros derivados de la matríz de confusión para evaluar el desempeño de estos clasificadores, nos sugieren los mismos criterios que nos bridan las curvas ROC de los clasificadores. En la Tabla 3.3.5.1.4 se muestran los valores de los parámetros mencionados.

Tabla 3.3.5.1.1. Funciones discriminantes canónicas obtenidas con la introducción de todos los aminoácidos que satisfacen el test de tolerancia y con el método Stepwise.

Aminoácidos	Función discriminante	
	Todas	Stepwise
Alanina	0.2269488	0.8519413
Cisteína	0.9328881	-
Ácido Aspártico	1.8950519	-0.852225
Ácido Glutámico	0.7074865	-
Fenilalanina	1.5171704	-
Glicina	0.6465456	-
Histidina	1.2819573	-
Isoleucina	0.077029	0.5565968
Lisina	0.7912264	-
Leucina	-0.291125	1.0783397
Metionina	3.6952632	-2.950158
Asparagina	1.6428562	-
Prolina	-0.516774	1.4921253
Glutamina	0.7912194	-
Arginina	1.2166805	-
Serina	1.6386476	-1.005287
Treonina	0.70664	-
Valina	0.852321	-
Tirosina	-	-
Triptófano	-0.898853	1.6170774
(Constant)	-47.3287	-5.956623

Tabla 3.3.5.1.2. Eficacia de las funciones discriminantes a través de las correlaciones canónicas y los valores de la Lambda de Wilk.

Función	Valor principal	% de Varianza	% Var. Acum.	Corr. Canónica	Función	Lambda de Wilks	Chi cuadrado	g.l.	Sig.
Stepwise									
1	2.199	100	100	0.829	1	0.313	162.796	8	0.000
Todas las variables									
1	2.411	100	100	0.841	1	0.293	165.020	19	0.000

Tabla 3.3.5.1.3. Resultado del área bajo la curva en los tres métodos utilizados.

Resultados del Análisis	Área	Error Estándar	Sig. Asintótica	Intervalo de confianza asintótico para el 95%	
				Límite superior	Límite inferior
Probabilidad Mamifero (Análisis CHAID)	0.949	0.017	0.000	0.915	0.982
Probabilidad Mamifero (Análisis Disc. Todas)	0.989	0.005	0.000	0.978	0.999
Probabilidad Mamifero (Análisis Disc. Stepwise)	0.985	0.007	0.000	0.970	0.999

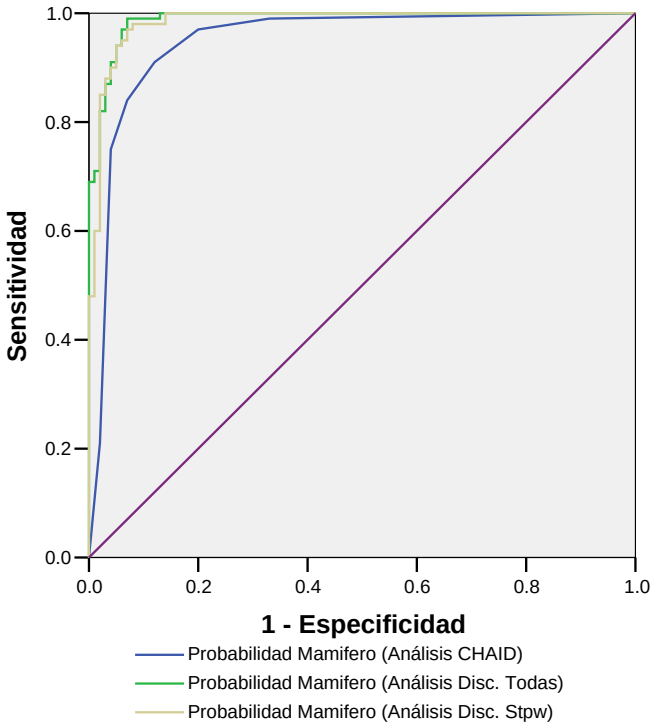


Figura 3.3.5.1.1 Curvas ROC obtenidas con los dos métodos de discriminante y con el método CHAID.

Tabla 3.3.5.1.4 Parámetros calculados a partir de la matriz de confusión para evaluar el desempeño de los clasificadores utilizados.

	Grupos	Razón de TP	Razón de TN	Prec.	Exac.	% de Clasf.
Predicciones de los miembros del Grupo con Anl. Discriminante (Stpw).						
70 % base de datos extendida	**Vert No Mamif**	91.7	94.6	94.3	93.2	91.7
	Mamiferos	94.6	91.7	92.1	93.2	94.6
Validación cruzada	**Vert No Mamif**	90.3	94.6	94.2	92.5	90.3
	Mamiferos	94.6	90.3	90.9	92.5	94.6
Validación externa	**Vert No Mamif**	96.4	96.2	96.4	96.3	96.4
	Mamiferos	96.2	96.4	96.2	96.3	96.2
Predicciones de los miembros del Grupo con Anl. Discriminante (Todas).						
70 % base de datos extendida	**Vert No Mamif**	93.1	93.2	93.1	93.2	93.1
	Mamiferos	93.2	93.1	93.2	93.2	93.2
Validación cruzada	**Vert No Mamif**	88.9	90.5	90.1	89.7	88.9
	Mamiferos	90.5	88.9	89.3	89.7	90.5
Validación externa	**Vert No Mamif**	96.4	100.0	100.0	98.1	96.4
	Mamiferos	100.0	96.4	96.3	98.1	100.0
Predicciones de los miembros del Grupo con CHAID						
70 % base de datos extendida	**Vert No Mamif**	91.7	93.2	93.0	92.5	91.7
	Mamiferos	93.2	91.7	92.0	92.5	93.2
Validación externa	**Vert No Mamif**	78.6	84.6	84.6	81.5	78.6
	Mamiferos	84.6	78.6	78.6	81.5	84.6

3.3.6. Aminoácidos asociados mediante el uso de codones con las clasificaciones taxonómicas en primates y homo sapiens.

Teniendo en cuenta las peculiaridades de estas dos especies por su cercania en el árbol filogenético universal, explicadas el la sección dedicada a esta misma taxa pero para el estudio de las secuencias de aminoácidos, los resultados obtenidos con la aplicación del método CHAID, son esperados desde el punto de vista biológico, pues existe una aceptada correlación entre todos los aminoácidos. Solamente tres de ellos no alcanzan valores significativos, la Serina, la Leucina y Cisteína, mientras la mayoria muestra indices de clasificación por encima de 90%, como muestra la Tabla 3.3.6.1. Cuando se realiza una validación al 70% de la base curada, Tabla 3.3.6.2, se observa que los porcientos de clasificación son inferiores a los obtenidos anteriormente, sin embargo es de esperar que con

una base externa con este método usando la probabilidad en el uso de codones se diferencien bien las especies involucradas en esta taxa, lo cual contribuye a la verificación de una de nuestras hipótesis de investigación.

Tabla 3.3.6.1. Significación de los aminoácidos al ser utilizados como variables predictoras en la construcción de árboles de decisión y los porcientos de clasificación alcanzados.

AA	Sig.	%clasificación
Glicina	0.024266804	93
Metionina	0.016568011	94
Histidina	0.015554361	92
Valina	0.003344738	89
Glutamina	0.002464145	88
Asparagina	0.002171647	85
Alanina	0.000387709	85
Lisina	0.000319111	91
Fenilalanina	0.000218361	86
Arginina	8.5694E-05	91
Prolina	2.31555E-05	91
Isoleucina	2.87011E-06	91
Triptófano	1.68E-06	89
Tirosina	8.32568E-07	92
Ácido Aspártico	2.73194E-07	94
Ácido Glutámico	5.53884E-08	93
Treonina	1.04E-08	92

[a]Sig. Significación del estadígrafo de razón verosimilitud Chi-cuadrado. Por simplificación se ha utilizado el simbolismo del SPSS para la notación científica, es decir, por ejemplo, el símbolo E-05 significa 10^{-5}.

Tabla 3.3.6.2. Clasificación obtenida con método CHAID en la base de datos tomando aleatoriamente el 70% de la base como entrenamiento y el resto usado en validación externa.

		Predicciones		
muestra	observada	primate	homoSapiens	Exactitud
entrenamiento	primate	36	3	92.3%
	homoSapiens	4	33	89.2%
	% Total	52.6%	47.4%	90.8%
valid	primate	7	4	63.6%
	homoSapiens	2	11	84.6%
	% Total	37.5%	62.5%	75.0%

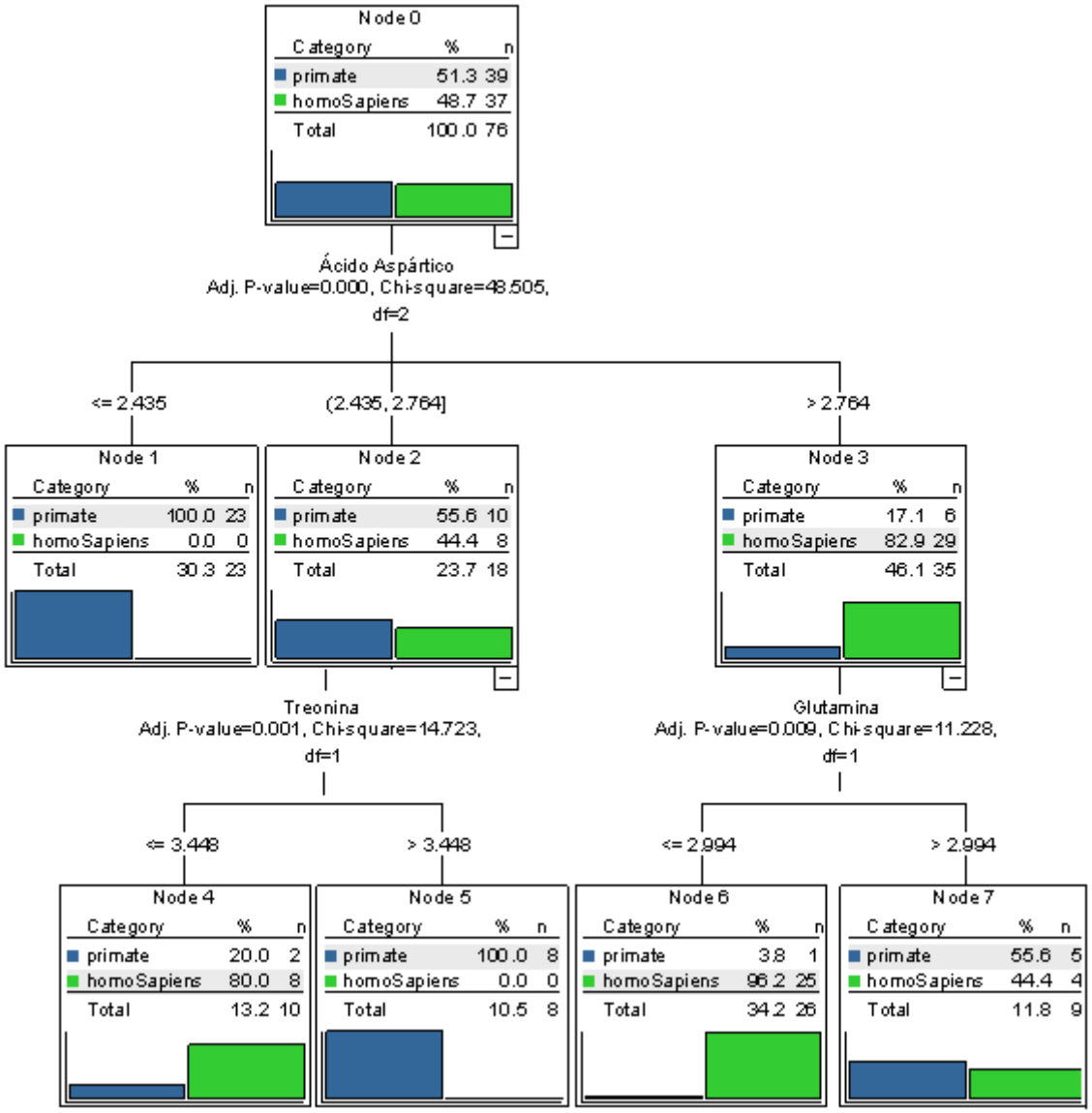


Figura 3.3.6.1 Arbol de Aminoácidos asociados con los resultados de una validación cruzada en la base curada con las clasificaciones taxonómica de primates y homo sapiens.

3.3.6.1. Análisis de Discriminante y la evaluación del desempeño de los clasificadores.

Con el análisis Discriminante realizado en esta taxa en cuanto al uso de codones se ratifica los resultados con el método CHAID, pues lo porcientos de clasificación mejoran considerablemente. Podemos observar en la Tabla 3.3.6.1.1 que las funciones discriminantes canónicas obtenidas con la introducción de todos los aminoácidos que

satisfacen el test de tolerancia, solo la Tirosina no aparece mientras que con el método Stepwise aparecen solamente el ácido Aspártico, la Treonina y la Tirosina.

En la Tabla 3.3.6.1.2 se puede apreciar que la eficacia de las funciones discriminantes en la separación de los casos en grupos, expresada a través de las correlaciones canónicas, es similar para ambos procedimientos. Además, los valores de la Lambda de Wilk y la significación del test Chi-cuadrado indican que las capacidades discriminatorias de las funciones obtenidas por estos procedimientos son similares, indicando el buen desempeño de las funciones discriminantes.

Con las curvas ROC se ilustra Figura 3.3.6.1.1 que el análisis Discriminante supera al CHAID, las difencias en los valores de las áreas bajo la curva Tabla 3.3.6.1.3, ratifican el hecho de que aunque las diferencias no son altamente significativas, el Intervalo de confianza asintótico para el 95% del CHAID queda completamente incluído en los intervalos de los métodos de Discriminante, mostrando su superioridad a la hora de la clasificación de estos organismos.

Al utilizar los parámetros derivados de la matríz de confusión para evaluar el desempeño de estos clasificadores, también nos sugieren que existen algunas diferencias entre ellos, en la Tabla 3.3.6.1.4 se muestran los valores de los parámetros mencionados.

Tabla 3.3.6.1.1. Funciones discriminantes canónicas obtenidas con la introducción de todos los aminoácidos que satisfacen el test de tolerancia y con el método Stepwise.

Aminoácidos	Función discriminante	
	Todas	Stepwise
Alanina	3.2271755	-
Cisteína	2.3901049	-
Ácido Aspártico	4.3812572	-2.830154
Ácido Glutámico	2.9292366	-
Fenilalanina	2.6504594	-
Glicina	2.0562512	-
Histidina	1.6416941	-
Isoleucina	2.9344641	-
Lisina	1.2136819	-
Leucina	2.4635822	-
Metionina	0.9738945	-
Asparagina	2.571989	-
Prolina	2.4680417	-
Glutamina	1.4374376	-
Arginina	2.3757739	-
Serina	1.889281	-
Treonina	-1.243473	3.2178151
Valina	0.7428819	-
Tirosina	-	2.2715964
Triptófano	1.4286537	-
(Constant)	-122.672	-7.493221

Tabla 3.3.6.1.2. Eficacia de las funciones discriminantes a través de las correlaciones canónicas y los valores de la Lambda de Wilk.

Función	Valor principal	% de Varianza	% Var. Acum.	Corr. Canónica	Función	Lambda de Wilks	Chi cuadrado	g.l.	Sig.
Stepwise									
1	2.827	100	100	0.859	1	0.261	97.297	3	0.000
Todas las variables									
1	4.062	100	100	0.896	1	0.198	104.601	19	0.000

Tabla 3.3.6.1.3. Resultado del área bajo la curva en los tres métodos utilizados.

Resultados del Análisis	Área	Error Estándar	Sig. Asintótica	Intervalo de confianza asintótico para el 95%	
				Límite superior	Límite inferior
Probabilidad Homo Sapiens (Análisis Disc. Stepwise)	0.999	0.002	0.000	0.996	1.002
Probabilidad Homo Sapiens (Análisis CHAID)	0.932	0.027	0.000	0.879	0.984
Probabilidad Homo Sapiens (Análisis Discriminante)	0.999	0.002	0.000	0.996	1.002

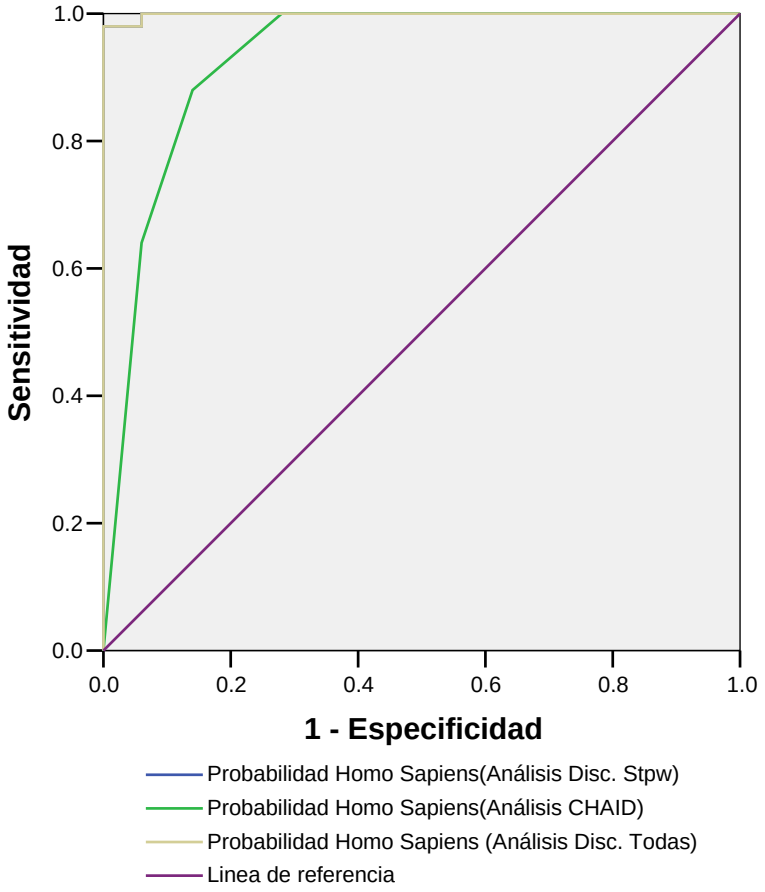


Figura 3.3.6.1.1 Curvas ROC obtenidas con los dos métodos de discriminante y con el método CHAID.

Tabla 3.3.6.1.4 Parámetros calculados a partir de la matriz de confusión para evaluar el desempeño de los clasificadores utilizados.

	Grupos	Razón de TP	Razón de TN	Prec.	Exac.	% de Clasf.
Predicciones de los miembros del Grupo con Anl. Discriminante (Stpw).						
70 % base de datos extendida	**Primates**	94.9	100.0	100.0	97.4	94.9
	HomoS	100.0	94.9	94.9	97.4	100.0
Validación cruzada	**Primates**	87.2	100.0	100.0	93.4	87.2
	HomoS	100.0	87.2	88.1	93.4	100.0
Validación externa	**Primates**	90.9	100.0	100.0	95.8	90.9
	HomoS	100.0	90.9	92.9	95.8	100.0
Predicciones de los miembros del Grupo con Anl. Discriminante (Todas).						
70 % base de datos extendida	**Primates**	97.4	97.3	97.4	97.4	97.4
	HomoS	97.3	97.4	97.3	97.4	97.3
Validación cruzada	**Primates**	84.6	94.6	94.3	89.5	84.6
	HomoS	94.6	84.6	85.4	89.5	94.6
Validación externa	**Primates**	100.0	100.0	100.0	100.0	100.0
	HomoS	100.0	100.0	100.0	100.0	100.0
Predicciones de los miembros del Grupo con CHAID						
70 % base de datos extendida	**Primates**	92.3	89.2	90.0	90.8	92.3
	HomoS	89.2	92.3	91.7	90.8	89.2
Validación externa	**Primates**	63.6	84.6	77.8	75.0	63.6
	HomoS	84.6	63.6	73.3	75.0	84.6

4. ANÁLISIS FILOGENÉTICOS.

La reconstrucción de la historia evolutiva de genes y especies es actualmente uno de los asuntos más importantes en la evolución molecular. En la medida en que los análisis filogenéticos realizados sean fiables, ellos verterán la luz en la sucesión de eventos evolutivos que han generado la diversidad de hoy día de las especies y nos ayuda a entender los mecanismos de evolución así como la historia de organismos.

4.1. ANÁLISIS FILOGENÉTICOS EN LA BASE DE PROTEINAS.

La filogenia es la ciencia de estimar el pasado, en particular la filogenia molecular basada en comparación de secuencias de proteínas o de DNA. Un árbol filogenético es un árbol que muestra las relaciones de evolución entre varias especies u otras entidades que se cree que tuvieron una descendencia común, además se consideran una estructura matemática que se usa para modelar la historia evolutiva de un grupo de secuencias o de organismos. Usa información proveniente de fósiles así como aquélla generada por la comparación estructural y molecular. En nuestro trababjo se comparan secuecias de organismos actuales de una base datos curados con la verificación en una base extendida descritas ambas en el Capitulo 2.

Los árboles filogenéticos se construyen tomando en cuenta la teoría de la evolución, que nos indica que todos los organismos son descendientes de un ancestro común: la protocélula ver anexo 1. Así, todos los organismos, ya sean vivos o extintos, se encuentran emparentados en algún grado.

Para la obtención de los árboles se utilizó el MEGA 4. En la sección 1.2.2 se explica lo relacionado con las posibilidades que este software brinda y las herramientas que fueron utilizadas en el trabajo con el mismo. En particular, el uso de este software nos permitió seleccionar una función de distancia apropiada entre los vectores NECk que nos permitiera obtener árboles plausibles desde el punto de vista evolutivo, los cuales no se encontraran en abierta contradicción con las observaciones y evidencias biológicas.

El uso del MEGA permitió verificar que si las bases de datos correspondientes a cada grupo taxonómico se sobrecargan con secuencias de proteínas vinculadas a procesos

biológicos esenciales para todas las células vivas entonces, al construir el árbol filogenético se obtienen ramas ubicadas de forma errónea en el árbol. Un ejemplo concreto se obtiene al sobrecargar la base de invertebrados con proteínas involucradas en las cadenas de transporte de electrones, un proceso esencial para todas las células vivas. En particular, la familia de los citocromos, vinculadas con estos procesos, se caracteriza por poseer dominios estructurales en sus secuencias de aminoácidos conservadas, en la mayoría de los taxa, desde los procariotes hasta el homo sapiens. Este hecho provoca que gran parte de la información estadística reflejada en los vectores NEC_k sea común para mayoría de los taxa. Como consecuencia se obtiene el efecto que se observa en el árbol de la Figura 4.1.1, en el cual los invertebrados (sin incluir los insectos) se ubican en una rama próxima a los primates, cuando, desde un punto de vista evolutivo, deben ubicarse en una rama contígua al ancestro de los vertebrados. Sin embargo, los insectos (invertebrados no incluidos en el taxa que lleva este nombre) respecto a los vertebrados se ubican en una rama con mayor sentido evolutivo. La causa de este resultado se explica debido a que la base de insectos posee un mejor balance en cuanto a la variabilidad de los tipos de proteínas. Debemos mencionar que estos efectos tienen lugar debido a la naturaleza estadística de la información utilizada, pues para construir los árboles las matríces de distancias no se calculan directamente de las secuencias de proteínas alineadas, como en el análisis filogenético clásico, sino que se estiman a partir de vectores que expresan regularidades estadísticas presentes en las secuencias no alineadas.

Finalmente, la construcción de las bases de datos teniendo en cuenta las restricciones biológicas descritas en el capítulo 2, permitió construir un árbol filogenético que muestra resultados importantes desde el punto de vista evolutivo en el reino animal [34] (Figuras 4.1.2 y 4.1.3).

Comúnmente cada árbol construido se valida en alguna medida utilzando un procedimiento bootstrap. En particular, cuando se parte de secuencias de proteínas alineadas y se utilizan las funciones de distancia que tiene por defecto el MEGA4, este software tiene la opción de construir 500 árboles y llegar a un árbol consenso que alcance el 70%. Tal procedimiento no es aplicable a nuestro caso. Luego, para obtener un árbol consenso se realizaron muestreos aleatorios de las bases de datos y a partir de cada submuestra generada se calcularon las matrices de distancia ver Anexo 9. Tomando una

selección aleatoria del 90% de la base se construyen las primeras 100 matrices obteniéndose un árbol consenso que representa el 72%. Al construir 100 más se obtiene un árbol consenso que representa el 80% de los doscientos posibles árboles. Lo anterior corrobora la información que brinda el árbol obtenido de nuestra base de datos, dando respuesta asi a nuestra segunda interrogante de investigación.

Este resultado, además de estar en correspondencia con el árbol filogenético evolutivamente esperado, presenta la peculiaridad que dos grupos de organismos, los vertebrados no mamiferos y los mamiferos están enraizados en el mismo nodo lo que sugiere una pérdida de información acerca de los ancestros de estos taxa, en algún momento del proceso evolutivo. La causa de esta pérdida de información pudo estar determinada por un proceso de extinción a gran escala, a partir del cual los grupos de organismos sobrevivientes, adaptados a un medio ambiente que les permitió sobrevir durante la extinción, eran portadores de caracteres genéticos comunes, los cuales pudieron ser frutos de un proceso de evolución convergente estimulado por el ecosistema en que se desarrollaron. A lo largo de la historia evolutiva de las especies hay varios ejemplos de este tipo de evolución (ver ejemplo en el sumario biológico, capítulo 1). No obstante en nuestra investigación nos dimos a la tarea de corroborar, en la literatura actualizada, primero la existencia de grupos de mamíferos desde la Era Mesozoica donde dominaban los vertebrados no mamíferos y segundo aquellos procesos de extinción que involucraron a estos organismos y lo que los caracterizó pudiendo referenciar criterios científicos como:

- El carácter fundamental de la Era Mesozoica, en cuanto a lo que a Vertebrados se refiere, es el desarrollo inusitado que durante esta época tuvieron los reptiles, adaptándose a diversos medios ecológicos tanto continentales (estegosauros, tyrannosaurus, triceratops), como marinos (plesiosaurios, ictiyosaurios, mosasaurios), y aéreos (pterosaurios), donde llegaron a desarrollar grandes dimensiones. Se inicia entonces el desarrollo de todo el conjunto de reptiles que llegan hasta la actualidad (cocodrilos, quelonios, saurios, ofidios), así como el grupo de los terápsidos, que son los **precursores de los mamíferos**. La mayoría de estos grupos aparecen entre el Pérmico y el Triásico, que son los periodos de máxima expansión reptiliana.

- Los restos más antiguos de **mamíferos**, proceden del Triásico superior. En el Mesozoico los fósiles de mamíferos son escasos, en general, grupos especiales de organismos no placentados (marsupiales). Los primeros registros fósiles de mamíferos placentados corresponden con materiales de finales del Cretácico (en Mongolia), que corresponden a organismos de pequeña talla, tipo de los Insectívoros y con caracteres muy primitivos.
- En las superficies continentales la mayor expansión corresponde a los mamíferos (presentes desde el Mesozoico). Los marsupiales desarrollan numerosas formas adaptativas en Australia y América del Sur, durante la ausencia de predadores carnívoros placentados, ya que la diversidad de los mamíferos placentados en general, presenta una evolución genética mucho más eficaz. En la actualidad más del 95% de los mamíferos conocidos son placentarios.

Con los argumentos anteriores se corrobora la existencia de mamíferos con características muy peculiares, por su forma de adaptación al medio, presentes en la época resplandeciente de los grandes reptiles. En particular, dos carácterísticas comunes a la mayoría de estos mamíferos es su pequeño tamaño y la presencia de adaptaciones que les permitían vivir bajo tierra en la salvaguarda de los grandes depredadores. Es bien conocido que todos los animales que se adaptan a un mismo ambiente, independientemente de la clase a la que pertenezcan desarrollan caracteres genéticos similares que les permiten sobrevivir en dicho medio ambiente. Por lo que se sugiere que antes de la ocurrencia de la gran extinción masiva pudo tener lugar la evolución convergente de muchos caracteres presentes en mamíferos y reptiles pequeños, las cuales le permitieron sobrevivir a la extinción. La evolución convergente de muchos de estos caracteres debió quedar grabada en las regularidades estadísticas encontradas en los genes y proteínas actuales derivadas de genes y proteínas de los mamiferos y reptiles ancestros que sobrevivieron al proceso de extinción. Este análisis explicaría la aparente presencia de un “ancestro común” entre vertebrados no mamíferos y mamíferos mostrada en la Figura 4.1.2A y la posible pérdida de información causada durante la extinción:

- Hacia finales del Mesozoico ocurrió una extinción masiva en el Cretáceo terciario. Este fue el evento de extinción que acabó con los dinosaurios (entre otros). Muchos de los animales y plantas que sobrevivieron (tales como **mamíferos y aves**) se multiplicaron después del Cenozoico. Los mamíferos, que eran pequeños y poco abundantes durante el Mesozoico, se hicieron más diversos. Nuevas especies de mamíferos evolucionaron y fueron capaces de vivir y alimentarse en áreas usadas por los dinosaurios durante el Mesozoico, según investigadores dirigidos por Olaf Bininda-Emonds, de la universidad Jena de Friedrich-Schiller, en Alemania.

Otra posibilidad de analizar la pérdida de información acerca de los ancestros de estos dos grupos de organismos es la exclusión de uno de ellos para ver el comportamiento del árbol. Cuando se excluye el grupo de vertebrados que no son mamíferos se obtiene un árbol en correspondencia con lo discutido aquí, luego se realiza la prueba de construir un determinado número de matrices de distancia con una selección aleatoria del 90% de la base extendida aleatorizada y con las primeras cien pruebas se obtiene un árbol consenso del 70 %. En el árbol filogenético mostrado en la Figura 4.1.2B se evidencia que debió existir un ancestro común de todos los mamíferos actuales. Esta evidencia proporciona una respuesta estadística a una de nuestras preguntas de investigación y corrobora lo sucedido en el proceso de evolución y su estrecha relación con las probabilidades de aparición de un aminoácido en una secuencia de proteínas.

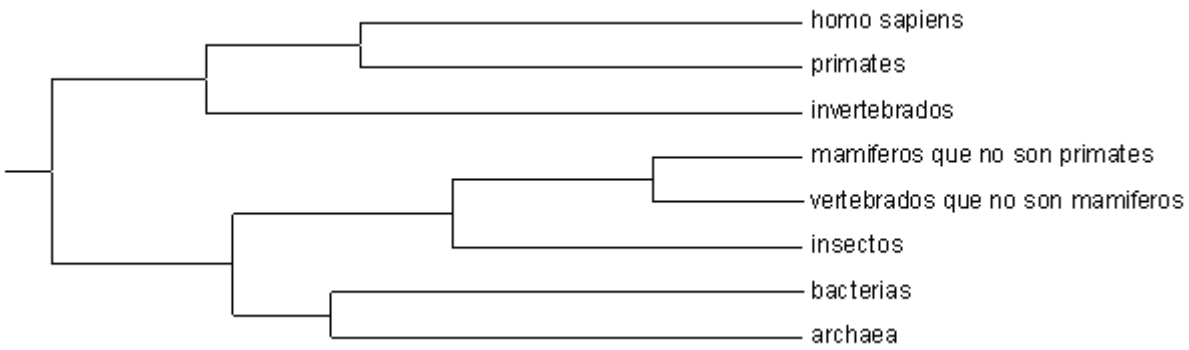


Figura 4.1.1. Árbol obtenido con base de datos donde el grupo de invertebrados tenía un por ciento considerable de proteínas del tipo Cytochrome (transporte) conservadas en el proceso de evolución.

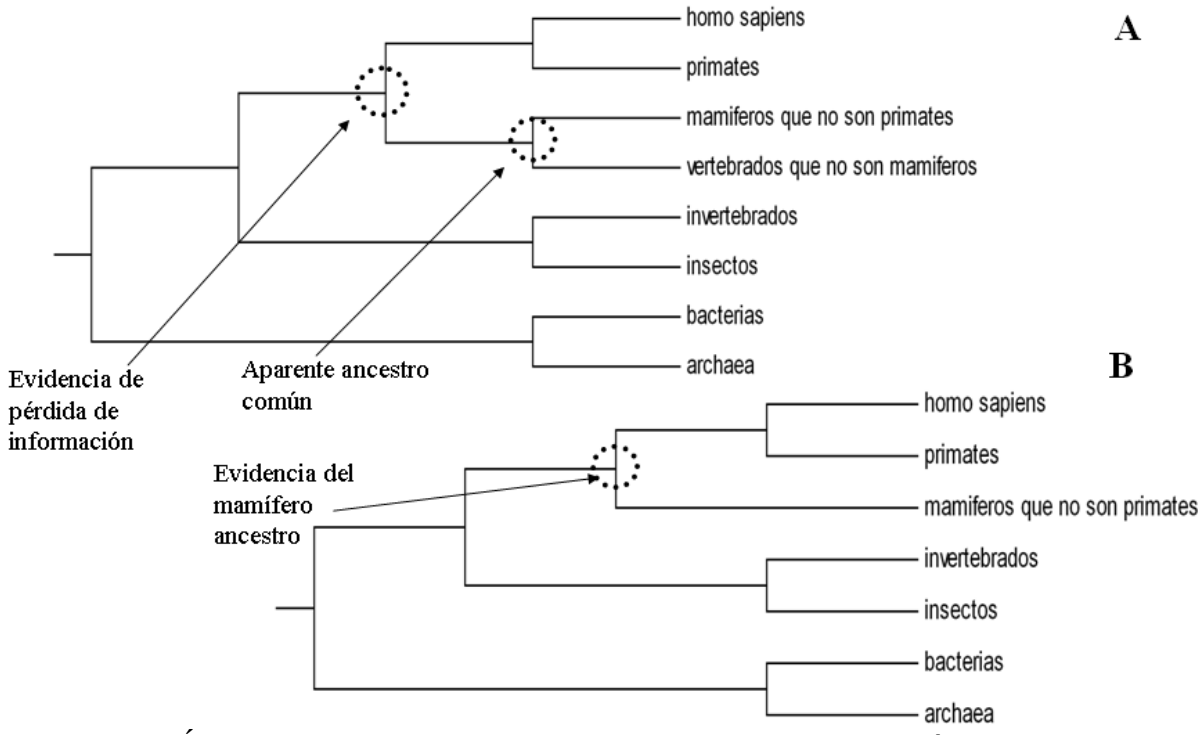


Figura 4.1.2. Árbol obtenido con base de datos curada. A: Logrando un árbol consenso del 80%, con la construcción de 200 matrices de la base de datos extendida. B: Verificando el hecho que excluyendo los vertebrados no mamíferos el comportamiento es el mismo y se obtiene un árbol consenso del 70 % con las primeras cien matrices de la base de datos extendida.

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

A partir de los resultados obtenidos podemos concluir que:

- Los análisis realizados con los vectores *NECk*, calculados a partir de las secuencias de proteínas y del uso de codones en los genes, nos permitieron detectar diferencias estadísticamente significativas entre los taxa estudiados en correspondencia con la clasificación taxonómica.
- Mediante el uso de la distancia de Hellinger entre los vectores estimados de distribución de probabilidades de aparición de aminoácidos en las proteínas, fue posible detectar relaciones filogenéticas entre los taxa estudiados en concordancia con la taxonomía evolutiva.

Recomendaciones

1. Realizar un análisis filogenético usando los vectores *NECk* calculados a partir de las bases uso de codones similar al realizado para los vectores *NECk* calculados a partir de las bases de secuencias de proteínas.
2. Investigar la variación de las distribuciones de las frecuencias de los aminoácidos en función del tiempo evolutivo transcurrido $\boxed{\frac{\partial p(x,t)}{\partial t} \approx \text{Constante}}$

REFERENCIAS BIBLIOGRÁFICAS

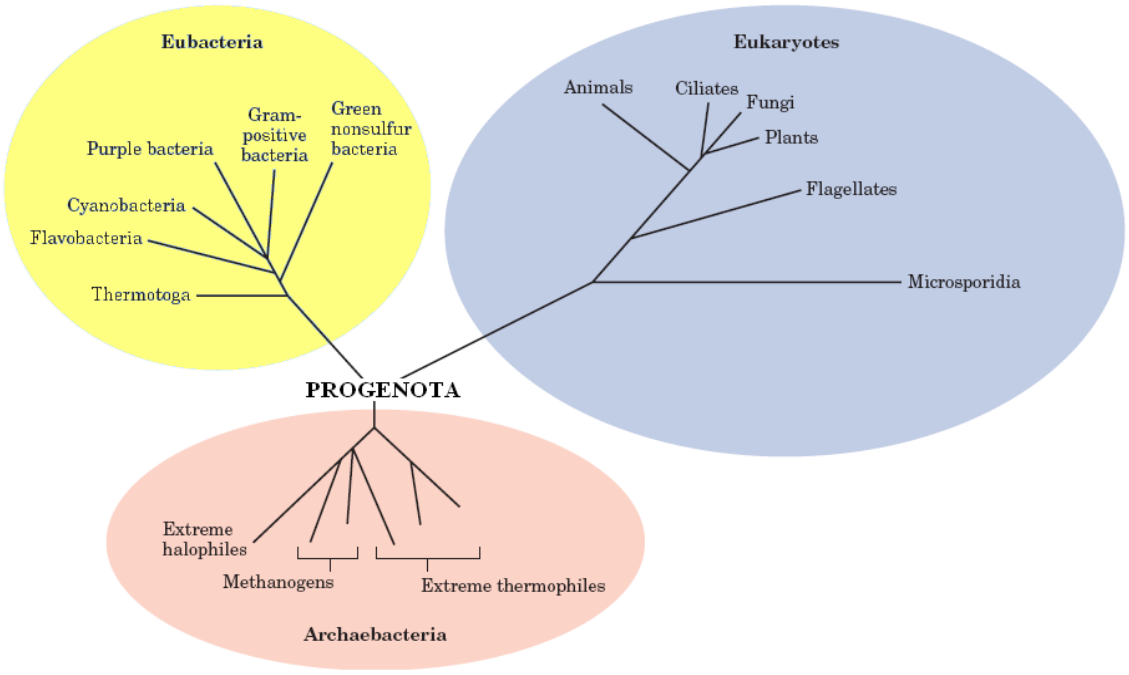
1. Lewin, B. Genes VIII. Pearson Prentice Hall. 2004.
2. Crick, F.H.C. The origin of the genetic code. J. Mol. Biol. 38, 367-379, 1968.
3. Knight RD, Freeland SJ, Landweber LF, 2001. Rewiting the keyboard: evolvability of the genetic code. Nat Rev Gente, 2:49-58.
4. Gillis, D; Massar, S.; Cerf, N.J. y Rooman, M. (2001) Optimality of the genetic code with respect to protein stability and amino acid frequencies. Genome Biology 2, research0049.1-research0049.12, 2001.
5. Epstein, C. J. Role of the amino-acid "code" and of selection for conformation in the evolution of proteins. Nature 210, 25-28, 1966
6. Epstein C. Non randomnes of amino-acid changes in the evolution of homologous proteins. Nature, 215, 355-359, 1967
7. Freeland, S. y Hurst, L. The genetic code is one in a million. J. Mol. Evol. 47, 238-248, 1998.
8. Frappat, L., Sciarrino A. y Sorba, P. "A crystal base for the genetic code" Phys. Lett. A250, 214-221, 1998.
9. Woese, C.R. On the evolution of the genetic code. Proc. Natl. Acad. Sci. USA 54, 1546-1552, 1965.
10. Haig, D. y Hurst, L. D. A quantitative measure of error minimization in the genetic code. J. Mol. Evol. 33, 412-417, 1991.
11. Friedman, S.M. y Weinstein, I.B. Lack of fidelity in the translation of ribopolynucleotides. Proc. Natl. Acad. Sci. USA, 52, 988-996, 1964
12. Parker J. Errors and alternatives in reading the universal genetic code. Microbiol. Rev. 53, 273-298, 1989.
13. Yang, Z.: Adaptive molecular evolution. In Handbook of statistical genetics, (Balding, M., Bishop, M. & Cannings, C., eds), Wiley:London, pp. 327-50, 2000.
14. Alff-Steinberger, C. The genetic code and error transmission. Proc. Natl. Acad. Sci. USA 64, 584-591, 1969

15. Nakamura Y, Gojobori T, y Ikemura T. Codon usage tabulated from international DNA sequence database: status for the year. Nucleic Acids Research 28, pp 292, 2000.
16. Makrides, S.C.: Strategies for achieving high-level expression of genes in Escherichia coli. Microbiol Rev 60, 512–38, 1996.
17. Duret, L., Mouchiroud, D.: Expression pattern and, surprisingly, gen length, shape codón usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci 96, 17–25, 1999.
18. Gu, W., Zhou, T., Ma, J., Sun, X., Lu, Z.: The relationship between synonymous codon usage and protein structure in Escherichia coli and Homo sapiens. Biosystems 73, 89-97, 2004.
19. Gupta, S.K., Majumdar, S., Bhattacharya, K., Ghosh, T.C.: Studies on the relationships between synonymous codon usage and protein secondary structure. Biochem Biophys Res Comm 269, 692-6, 2000.
20. Oresic. M., Shalloway, D.: Specific correlations between relative synonymous codon usage and protein secondary structure. J Mol. Biol. 281, 31–48, 1998.
21. Tao, X., Dafu, D.: The relationship between synonymous codon usage and protein structure. FEBS Lett 434, 93–6, 1998.
22. Fuglsang, A.: Strong associations between gene function and codon usage. APMIS 111, 843–7, 2003.
23. Sanchez, R.:“Estudio del orden en el Código Genético mediante la aplicación de métodos algebraico y estadístico, 2003.
24. Sanchez, R.:“Regularidades algebraicas del código genético: aplicaciones a la evolución molecular”. Tesis presentada en opción al grado científico de Doctor en Ciencias Biológicas, 2006.
25. Fawcett, T.: ROC Graphs: Notes and Practical Considerations for Data Mining Researchers, Hewlett-Packard Company, 2003.
26. Weiss, G. M., and Provost, F.: Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction, JAIR 19, 315–354, 2003.
27. University of Waterloo, Department of Statistics and Actuarial Science, SPSS Instruction Manual, September 1, 1998

28. Swanson, R. A unifying concept for the amino acid code. Bull. Math. Biol. 46, 187-203, 1984.
29. Gillis, D; Massar, S.; Cerf, N.J. y Rooman, M. (2001) Optimality of the genetic code with respect to protein stability and amino acid frequencies. Genome Biology 2, research0049.1-research0049.12, 2001.
30. Taylor, J.D.T. y Thornton, J.M. Recompilation of the mutation matrices. CABIOS 8, 275-282, 1991.
31. Kira S. Makarova, Yuri I. Wolf, Sergey L. Mekhedov, Boris G. Mirkin1 and Eugene V. Koonin. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. 4626–4638 Nucleic Acids Research, 2005, Vol. 33, No. 14
32. Koichiro Tamura, Joel Dudley, Masatoshi Nei, Sudhir Kumar. Center of Evolutionary Functional Genomics, Biodesign Institute, Arizona State University. MEGA Molecular Evolutionary Genetics Analysis. VERSION 4, 1993 - 2008.
33. CHAID, W. (1994). "CHAID para SPSS sobre Windows. Técnicas de segmentación basadas en razones de verosimilitud Chi-cuadrado." User Manual. SPSS Inc.
34. PhD Mohammad Badii1, Dr. Jerónimo Landeros2, Dr. Victoriano Garza3. Historia evolutiva de la vida, CULCyT//Enero –Febrero, 2008, Año 5, No 24
35. Ana Aber, Coordinadora, Alfredo Langguth, Editor, BIODIVERSIDAD Y TAXONOMÍA PRESENTE Y FUTURO. Resultados del Taller realizado en la Facultad de Ciencias, Universidad de la República. 14 - 18 de junio de 2004

ANEXOS

Anexos 1. Árbol Filogenético Universal.



Anexo 2. Fragmento de base de datos de cadenas de proteínas.

```
>gi|127069|sp|P16455|MGMT_HUMAN Methylated-DNA--protein-cysteine
methyltransferase (6-O-methylguanine-DNA methyltransferase) (MGMT) (O-6-
methylguanine-DNA-alkyltransferase)
MDKDCEMKRTTLDSPLGKLELSGCEQGLHEIKLLGKGTSAADAVEVPAPAAVLGGPEPLMQCTAWLNAYF
HQPEAIEEFPVPALHHPVFQQESFTRQVLWKLLKVVKFGEVISYQQLAALAGNPKAARAVGGAMRGNPVP
ILIPCHRVVCSSGAVGNYSGGLAVKEWLLAHEGHRLGKPGLGGSSGLAGAWLKGAGATSGSPPAGRN

>gi|74720969|sp|Q9UJV8|PURG_HUMAN Purine-rich element-binding protein
gamma
MERARRRGGGGGRGRGGKNVGGSGLSKSRLYPQAQHSHYPHYAASATPNQAGGAAEIQELASKRVDIQKK
RFYLDVKQSSRGRFLKIAEVWIGRGRQDNIRKSKLTLSLSVAAELKDCLGDFIEHYAHLGLKGHRQEHGH
SKEQGSRRRQKHSAPSPPVSVGSEEHPHSVLKTDYIERDNRKYYLDLKENQRGRFLRIRQTMMRGTGMIG
YFGHSLGQEQTIVLPAQGMIEFRDALVQLIEDYGEGDIEERRGGDDDPLELPEGTSFRVDNKRFYFDVGS
NKYGIFLKVSEVRPPYRNTITVPFKAWTRFGENFIKYEEEMRKICNSHKEKRMDGRKASGEEQECLD

>gi|1346918|sp|Q00577|PURA_HUMAN Transcriptional activator protein Pur-
alpha (Purine-rich single-stranded DNA-binding protein alpha)
MADRDSGSEQGGAALGSGGSLGHPGSGSGSGGGGGGGGGGGGSGGGGGGAPGGLQHETQELASKRVDIQN
KRFYLDVKQNAKGRFLKIAEVGAGGNKSRLTLSMSVAVEFRDYLGDFIEHYAQLGPSQPPDLAQAQDEPR
RALKSEFLVRENRKYYMDLKENQRGRFLRIRQTVNRGPGLGSTQGQTIALPAQGLIEFRDALAKLIDDYG
VEEEPAELPEGTSLTVDNKRFFFDVGSNKYGVFMRVSEVKPTYRNSITVPYKVWAKFGHTFCKYSEEMKK
IQEKQREKRAACEQLHQQQQQQQEETAAATLLLQGEEEGEED

>gi|13629600|sp|Q9Y2U8|MAN1_HUMAN Inner nuclear membrane protein Man1
(LEM domain-containing protein 3)mamifero
MAAAAASAPQQLSDEELFSQLRRYGLSPGPVTESTRPVYLKKLKKLREEEQQQHRSGGRGNKTRNSNNNN
TAAATVAAAGPAAAAAAGMGVRPVSGDLSYLRTPGGLCRISASGPESLLGGPGGASAAPAAGSKVLLGFS
SDESDVEASPRDQAGGGGRKDRASLQYRGLKAPPAPLAASEVTNSNSAERRKPHSWWGARRPAGPELQTP
PGKDGAVEDEEGEGEDGEERDPETEEPLWASRTVNGSRLVPYSCRENYSDSEEEDDDDVASSRQVLKDDS
LSRHRPRRTHSKPLPPLTAKSAGGRLETSVQGGGGLAMNDRAAAAGSLDRSRNLEEAAAAEQGGGCDQVD
SSPVPRYRVNAKKLTPLLPPPLTDMDSTLDSSTGSLLKTNNHIGGGAFSVDSPRIYSNSLPPSAAVAASS
SLRINHANHTGSNHTYLKNTYNKPKLSEPEEELLQQFKREEVSPTGSFSAHYLSMFLLTAACLFFLILGL
TYLGMRGTGVSEDGELSIENPFGETFGKIQESEKTLMMNTLYKLHDRLAQLAGDHECGSSSQRTLSVQEA
AAYLKDLGPEYEGIFNTSLQWILENGKDVGIRCVGFGPEEELTNITDVQFLQSTRPLMSFWCRFRRAFVT
VTHRLLLLCLGVVMVCVVLRYMKYRWTKEEEETRQMYDMVVKIIDVLRSHNEACQENKDLQPYMPIPHVR
DSLIQPHDRKKMKKVWDRAVDFLAANESRVRTETRRIGGADFLVWRWIQPSASCDKILVIPSKVWQGQAF
HLDRRNSPPNSLTPCLKIRNMFDPVMEIGDQWHLAIQEAILEKCSDNDGIVHIAVDKNSREGCVYVKCLS
PEYAGKAFKALHGSWFDGKLVTVKYLRLDRYHHRFPQALTSNTPLKPSNKHMNSMSHLRLRTGLTNSQGS
S

>gi|8475983|sp|O75916|RGS9_HUMAN Regulator of G-protein signaling 9
(RGS9)
MTIRHQGQQYRPRMAFLQKIEALVKDMQNPETGVRMQNQRVLVTSVPHAMTGSDVLQWIVQRLWISSLEA
QNLGNFIVRYGYIYPLQDPKNLILKPDGSLYRFQTPYFWPTQQWPAEDTDYAIYLAKRNIKKKGILEEYE
KENYNFLNQKMNYKWDFVIMQAKEQYRAGKERNKADRYALDCQEKAYWLVHRCPPGMDNVLDYGLDRVTN
PNEVKVNQKQTVVAVKKEIMYYQQALMRSTVKSSVSLGGIVKYSEQFSSNDAIMSGCLPSNPWITDDTQF
WDLNAKLVEIPTKMRVERWAFNFSELIRDPKGRQSFQYFLKKEFSGENLGFWEACEDLKYGDQSKVKEKA
EEIYKLFLAPGARRWINIDGKTMDITVKGLKHPHRYVLDAAQTHIYMLMKKDSYARYLKSPIYKDMLAKA
IEPQETTKKSSTLPFMRRHLRSSPSPVILRQLEEEAKAREAANTVDITQPGQHMAPSPHLTVYTGTCMPP
SPSSPFSSSCRSPRKPFASPSRFIRRPSTTICPSPIRVALESSSGLEQKGECSGSMAPRGPSVTESSEAS
LDTSWPRSRPRAPPKARMALSFSRFLRRGCLASPVFARLSPKCPAVSHGRVQPLGDVGQQLPRLKSKRVA
NFFQIKMDVPTGSGTCLMDSEDAGTGESGDRATEKEVICPWESL
```

Anexo 3. Fragmento de base de datos de uso de codones.

>AB000095\AB000095\176..1717\1542\BAA25014.1\Homo, sapiens\Homo, sapiens, mRNA, for, hepatocyte, growth, factor, activator, inhibitor,complete, cds./codon_start=1/product="hepatocyte, growth, factor, activator, inhibitor"/protein_id="BAA25014.1"/db_xref="GI:2924601"
0, 16, 8, 0, 1, 7, 3, 12, 19, 2, 0, 5, 1, 9, 1, 5, 8, 4, 6, 18, 8, 4, 8, 20, 3, 5, 3, 23, 5, 3, 7, 20, 7, 6, 3, 9, 22, 1, 3, 15, 22, 1, 2, 18, 10, 2, 13, 19, 23, 4, 12, 5, 25, 6, 17, 6, 1, 13, 3, 4, 7, 0, 0, 1

>AB000099\AB000099\106..462\357\BAA25877.1\Homo, sapiens\Homo, sapiens, mRNA, for, DCRB,, complete, cds./codon_start=1/product="DCRB"/protein_id="BAA25877.1"/db_xref="GI:3090432"
0, 0, 1, 0, 4, 1, 0, 3, 4, 3, 1, 2, 4, 2, 1, 4, 4, 2, 1, 3, 2, 1, 4, 2, 1, 5, 3, 3, 0, 3, 1, 0, 3, 0, 0, 1, 0, 1, 5, 2, 1, 0, 2, 2, 3, 1, 3, 1, 2, 6, 2, 0, 2, 1, 1, 2, 3, 4, 2, 1, 2, 0, 0, 1

>AB000114\AB000114\101..1366\1266\BAA19055.1\Homo, sapiens\Homo, sapiens, mRNA, for, osteomodulin,, complete, cds./codon_start=1/product="osteomodulin"/protein_id="BAA19055.1"/db_xref="GI:1769800"
1, 1, 0, 2, 5, 1, 13, 8, 4, 15, 6, 4, 8, 1, 0, 8, 4, 5, 6, 1, 0, 9, 15, 2, 1, 9, 4, 1, 0, 4, 5, 1, 2, 6, 4, 2, 4, 4, 18, 6, 9, 24, 17, 6, 9, 12, 26, 4, 10, 16, 10, 15, 4, 5, 12, 13, 6, 6, 14, 12, 1, 0, 1, 0

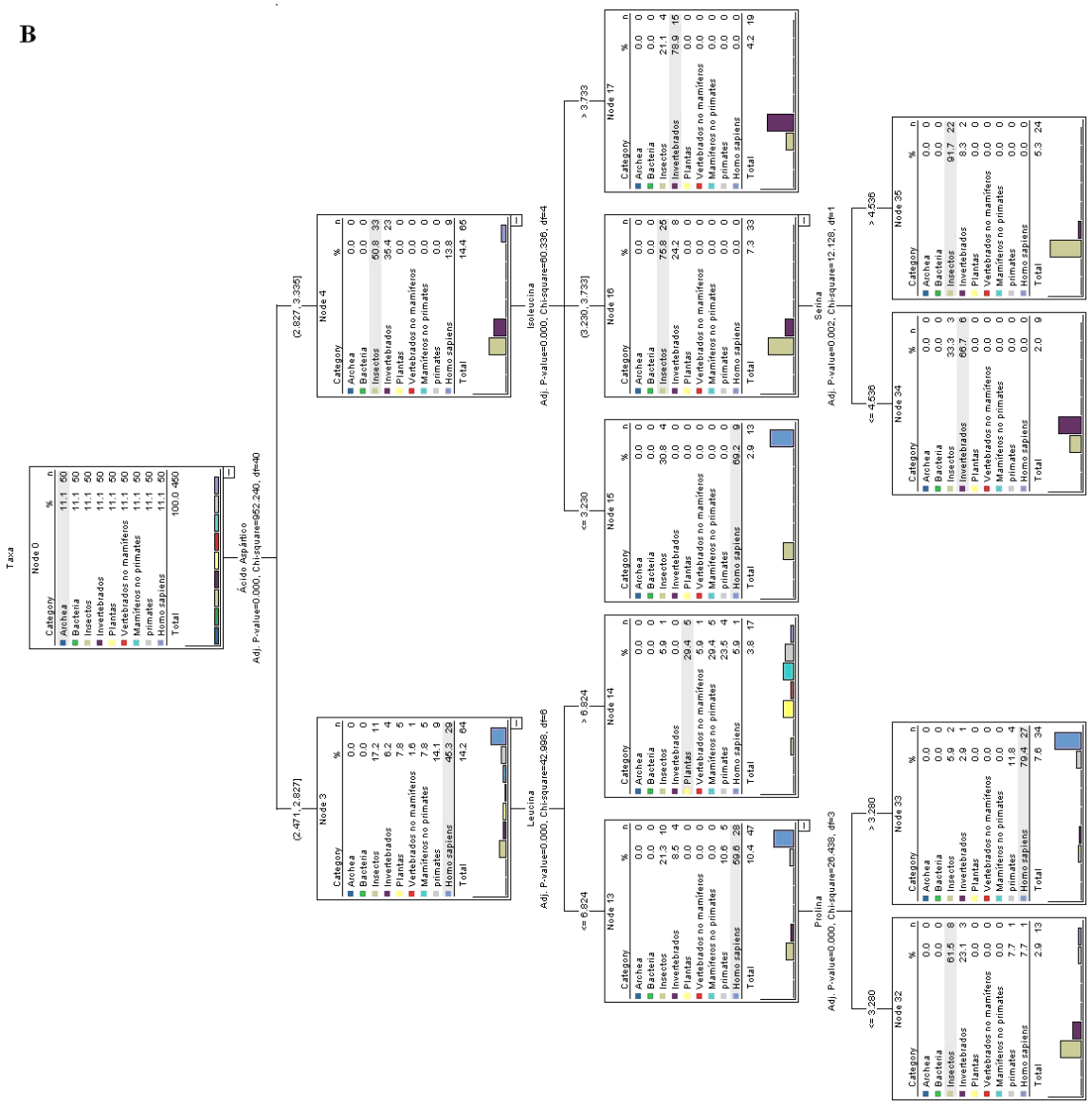
>AB000115\AB000115\242..1483\1242\BAA19056.1\Homo, sapiens\Homo, sapiens, mRNA, expressed, in, osteoblast,, complete, cds./codon_start=1/protein_id="BAA19056.1"/db_xref="GI:1769802"
2, 0, 3, 6, 8, 7, 7, 4, 7, 7, 10, 10, 4, 4, 1, 14, 3, 4, 5, 5, 1, 9, 7, 4, 0, 3, 10, 4, 1, 4, 9, 4, 4, 4, 3, 4, 8, 9, 17, 8, 6, 15, 6, 5, 5, 4, 13, 8, 14, 18, 3, 12, 4, 8, 2, 11, 9, 5, 23, 17, 1, 0, 0, 1

>AB000220\AB000220\563..2818\2256\BAA32398.1\Homo, sapiens\Homo, sapiens, mRNA, for, semaphorin, E,, complete, cds./codon_start=1/product="semaphorin, E"/protein_id="BAA32398.1"/db_xref="GI:3426163"
7, 5, 8, 3, 13, 13, 3, 5, 17, 7, 12, 7, 10, 13, 1, 17, 8, 11, 19, 9, 3, 19, 11, 7, 4, 12, 9, 9, 3, 16, 18, 10, 11, 5, 6, 10, 22, 16, 27, 19, 18, 22, 12, 20, 11, 13, 23, 11, 17, 26, 12, 14, 11, 12, 20, 16, 10, 12, 21, 17, 8, 1, 0, 0

>AB000221\AB000221\64..333\270\BAA21670.1\Homo, sapiens\Homo, sapiens, mRNA, for, CC, chemokine,, complete, cds./gene="PARC"/codon_start=1/product="CC, chemokine"/protein_id="BAA21670.1"/db_xref="GI:2289719"
0, 0, 1, 0, 1, 0, 1, 6, 2, 2, 0, 0, 0, 2, 0, 1, 2, 0, 0, 5, 0, 0, 2, 3, 0, 0, 2, 3, 0, 2, 0, 2, 0, 2, 0, 5, 0, 2, 2, 7, 1, 2, 2, 4, 0, 0, 1, 1, 3, 0, 1, 2, 5, 2, 1, 0, 1, 3, 1, 2, 2, 0, 0, 1

Anexo 4. Secciones B y C del árbol de aminoácidos asociados con las clasificaciones taxonómicas de organismos vivos

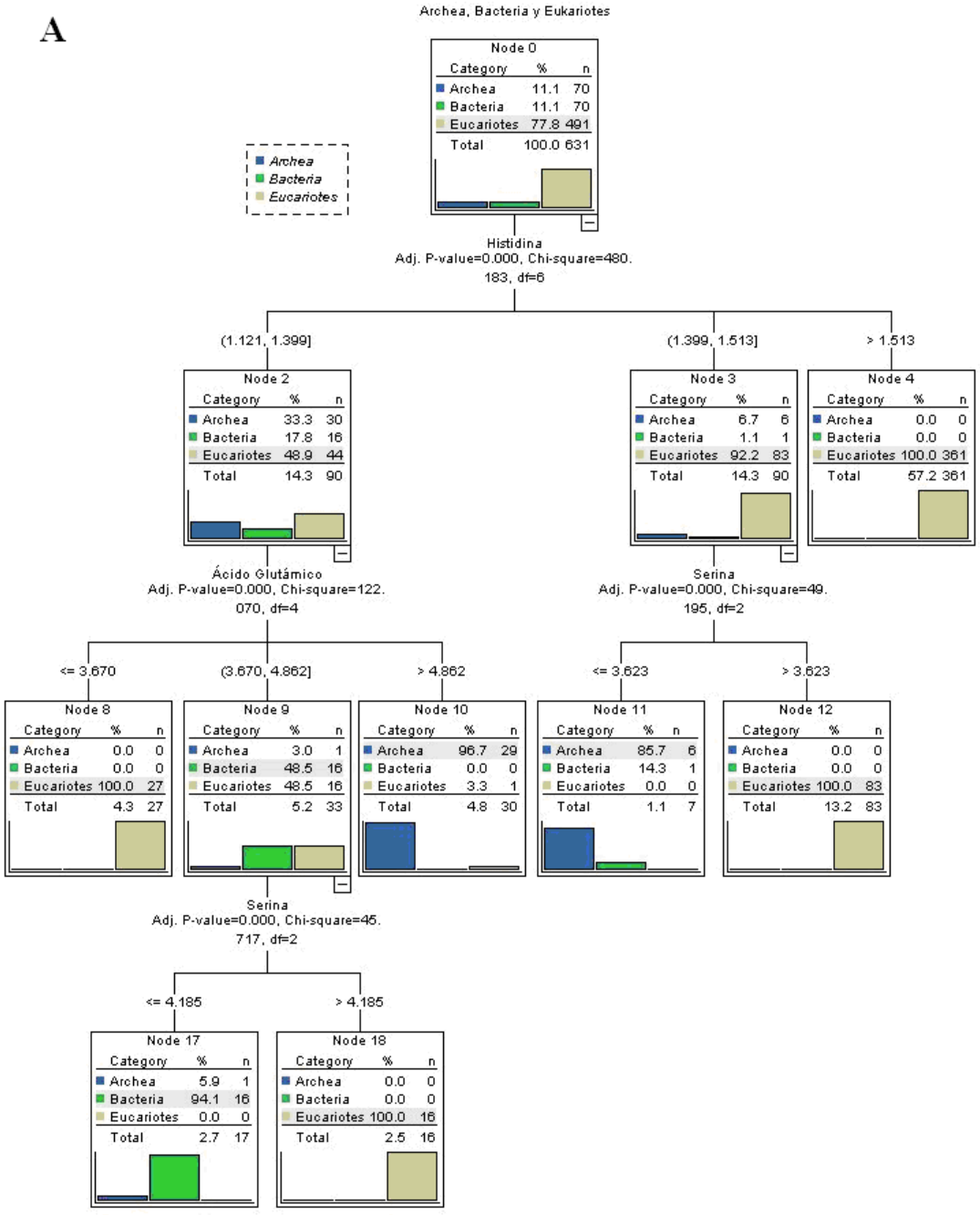
B



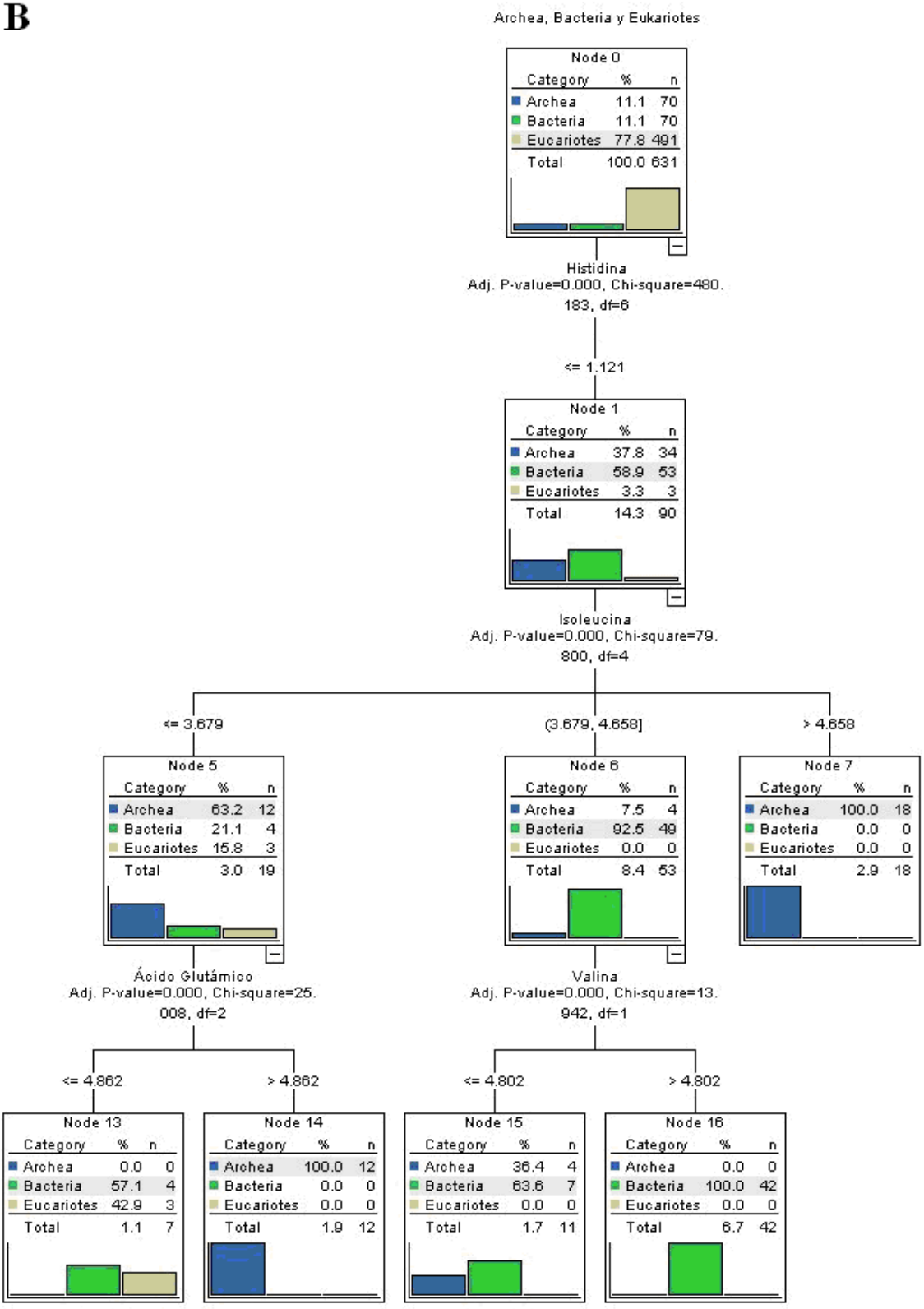

Taxa
Node 0
Category % n
Archea 11.1 50
Bacteria 11.1 50
Insectos 11.1 50
Invertebrados 11.1 50
Plantas 11.1 50
Vertebrados no mamíferos 11.1 50
Mamíferos no primates 11.1 50
primates 11.1 50
Homo sapiens 11.1 50
Total 100.0 450
Ácido Aspártico
Adj. P-value=0.000, Chi-square=952.240, df=40
<= 2.201
(2.201, 2.471]
Node 1
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 0.0 0
Plantas 26.4 34
Vertebrados no mamíferos 31.8 41
Mamíferos no primates 26.4 34
primates 14.7 19
Homo sapiens 0.8 1
Total 28.7 129
Node 2
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 1.6 1
Plantas 17.2 11
Vertebrados no mamíferos 12.5 8
Mamíferos no primates 17.2 11
primates 34.4 22
Homo sapiens 17.2 11
Total 14.2 64
Metionina
Adj. P-value=0.000, Chi-square=72.939, df=8
Metionina
Adj. P-value=0.000, Chi-square=59.413, df=10
<= 2.182
(2.182, 2.573]
> 2.573
<= 1.876
(1.876, 2.182]
> 2.182
Node 7
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 0.0 0
Plantas 11.1 2
Vertebrados no mamíferos 0.0 0
Mamíferos no primates 11.1 2
primates 72.2 13
Homo sapiens 5.6 1
Total 4.0 18
Node 8
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 0.0 0
Plantas 34.0 16
Vertebrados no mamíferos 21.3 10
Mamíferos no primates 34.0 16
primates 10.6 5
Homo sapiens 0.0 0
Total 10.4 47
Node 9
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 0.0 0
Plantas 25.0 16
Vertebrados no mamíferos 48.4 31
Mamíferos no primates 25.0 16
primates 1.6 1
Homo sapiens 0.0 0
Total 14.2 64
Node 10
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 0.0 0
Plantas 6.7 2
Vertebrados no mamíferos 0.0 0
Mamíferos no primates 6.7 2
primates 63.3 19
Homo sapiens 23.3 7
Total 6.7 30
Node 11
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 4.8 1
Plantas 33.3 7
Vertebrados no mamíferos 0.0 0
Mamíferos no primates 33.3 7
primates 14.3 3
Homo sapiens 14.3 3
Total 4.7 21
Node 12
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 0.0 0
Plantas 15.4 2
Vertebrados no mamíferos 61.5 8
Mamíferos no primates 15.4 2
primates 0.0 0
Homo sapiens 7.7 1
Total 2.9 13
Glicina
Adj. P-value=0.000, Chi-square=37.987, df=6
Tirosina
Adj. P-value=0.000, Chi-square=25.728, df=3
Leucina
Adj. P-value=0.000, Chi-square=21.479, df=3
Glutamina
Adj. P-value=0.002, Chi-square=21.000, df=4
<= 3.673
(3.673, 4.043]
> 4.043
<= 2.114
> 2.114
<= 6.824
> 6.824
<= 2.032
> 2.032
Node 23
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 0.0 0
Plantas 11.1 1
Vertebrados no mamíferos 22.2 2
Mamíferos no primates 11.1 1
primates 55.6 5
Homo sapiens 0.0 0
Total 2.0 9
Node 24
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 0.0 0
Plantas 45.2 14
Vertebrados no mamíferos 9.7 3
Mamíferos no primates 45.2 14
primates 0.0 0
Homo sapiens 0.0 0
Total 6.9 31
Node 25
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 0.0 0
Plantas 14.3 1
Vertebrados no mamíferos 71.4 5
Mamíferos no primates 14.3 1
primates 0.0 0
Homo sapiens 0.0 0
Total 1.6 7
Node 26
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 0.0 0
Plantas 11.4 4
Vertebrados no mamíferos 77.1 27
Mamíferos no primates 11.4 4
primates 0.0 0
Homo sapiens 0.0 0
Total 7.8 35
Node 27
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 0.0 0
Plantas 41.4 12
Vertebrados no mamíferos 13.8 4
Mamíferos no primates 41.4 12
primates 3.4 1
Homo sapiens 0.0 0
Total 8.4 29
Node 28
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 0.0 0
Plantas 0.0 0
Vertebrados no mamíferos 0.0 0
Mamíferos no primates 0.0 0
primates 22.2 2
Homo sapiens 77.8 7
Total 2.0 9
Node 29
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 0.0 0
Plantas 9.5 2
Vertebrados no mamíferos 0.0 0
Mamíferos no primates 9.5 2
primates 81.0 17
Homo sapiens 0.0 0
Total 4.7 21
Node 30
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 0.0 0
Plantas 50.0 7
Vertebrados no mamíferos 0.0 0
Mamíferos no primates 50.0 7
primates 0.0 0
Homo sapiens 0.0 0
Total 3.1 14
Node 31
Category % n
Archea 0.0 0
Bacteria 0.0 0
Insectos 0.0 0
Invertebrados 14.3 1
Plantas 0.0 0
Vertebrados no mamíferos 0.0 0
Mamíferos no primates 0.0 0
primates 42.9 3
Homo sapiens 42.9 3
Total 1.6 7

Anexo 5. Secciones A y B árbol y regla de clasificación de aminoácidos asociados con los resultados en la base de datos curada con validación cruzada en las clasificaciones taxonómicas de archaea, bacterias y eucariotes.

A



B


Archea, Bacteria y Eukariotes
Node 0
Category % n
Archea 11.1 70
Bacteria 11.1 70
Eucariotes 77.8 491
Total 100.0 631
Histidina
Adj. P-value=0.000, Chi-square=480.183, df=6
<= 1.121
Node 1
Category % n
Archea 37.8 34
Bacteria 58.9 53
Eucariotes 3.3 3
Total 14.3 90
Isoleucina
Adj. P-value=0.000, Chi-square=79.800, df=4
<= 3.679
(3.679, 4.658]
> 4.658
Node 5
Category % n
Archea 63.2 12
Bacteria 21.1 4
Eucariotes 15.8 3
Total 3.0 19
Node 6
Category % n
Archea 7.5 4
Bacteria 92.5 49
Eucariotes 0.0 0
Total 8.4 53
Node 7
Category % n
Archea 100.0 18
Bacteria 0.0 0
Eucariotes 0.0 0
Total 2.9 18
Ácido Glutámico
Adj. P-value=0.000, Chi-square=25.008, df=2
Valina
Adj. P-value=0.000, Chi-square=13.942, df=1
<= 4.862
> 4.862
<= 4.802
> 4.802
Node 13
Category % n
Archea 0.0 0
Bacteria 57.1 4
Eucariotes 42.9 3
Total 1.1 7
Node 14
Category % n
Archea 100.0 12
Bacteria 0.0 0
Eucariotes 0.0 0
Total 1.9 12
Node 15
Category % n
Archea 36.4 4
Bacteria 63.6 7
Eucariotes 0.0 0
Total 1.7 11
Node 16
Category % n
Archea 0.0 0
Bacteria 100.0 42
Eucariotes 0.0 0
Total 6.7 42

Regla de Clasificación

/* Node 13 */
IF (Histidina NOT MISSING AND (Histidina <= 1.12076082557669)) AND (Isoleucina NOT MISSING AND (Isoleucina <= 3.6787991498406)) AND (Ácido Glutámico NOT MISSING AND (Ácido Glutámico <= 4.86217846935535))
THEN
Node = 13
Prediction = 2
Probability = 0.571429

/* Node 14 */
IF (Histidina NOT MISSING AND (Histidina <= 1.12076082557669)) AND (Isoleucina NOT MISSING AND (Isoleucina <= 3.6787991498406)) AND (Ácido Glutámico IS MISSING OR (Ácido Glutámico > 4.86217846935535))
THEN
Node = 14
Prediction = 1
Probability = 1.000000

/* Node 15 */
IF (Histidina NOT MISSING AND (Histidina <= 1.12076082557669)) AND (Isoleucina IS MISSING OR (Isoleucina > 3.6787991498406 AND Isoleucina <= 4.65842040565458)) AND (Valina NOT MISSING AND (Valina <= 4.80227023068473))
THEN
Node = 15
Prediction = 2
Probability = 0.636364

/* Node 16 */
IF (Histidina NOT MISSING AND (Histidina <= 1.12076082557669)) AND (Isoleucina IS MISSING OR (Isoleucina > 3.6787991498406 AND Isoleucina <= 4.65842040565458)) AND (Valina IS MISSING OR (Valina > 4.80227023068473))
THEN
Node = 16
Prediction = 2
Probability = 1.000000

/* Node 7 */
IF (Histidina NOT MISSING AND (Histidina <= 1.12076082557669)) AND (Isoleucina NOT MISSING AND (Isoleucina > 4.65842040565458))
THEN
Node = 7
Prediction = 1
Probability = 1.000000

```
/* Node 8 */
IF (Histidina NOT MISSING AND (Histidina > 1.12076082557669 AND Histidina <=
1.39913310456926)) AND (Ácido Glutámico NOT MISSING AND (Ácido Glutámico <=
3.66998451669985))
THEN
        Node = 8
        Prediction = 3
        Probability = 1.000000

/* Node 17 */
IF (Histidina NOT MISSING AND (Histidina > 1.12076082557669 AND Histidina <=
1.39913310456926)) AND (Ácido Glutámico IS MISSING OR (Ácido Glutámico >
3.66998451669985 AND Ácido Glutámico <= 4.86217846935535)) AND (Serina IS MISSING OR
(Serina <= 4.18460680423871))
THEN
        Node = 17
        Prediction = 2
        Probability = 0.941176

/* Node 18 */
IF (Histidina NOT MISSING AND (Histidina > 1.12076082557669 AND Histidina <=
1.39913310456926)) AND (Ácido Glutámico IS MISSING OR (Ácido Glutámico >
3.66998451669985 AND Ácido Glutámico <= 4.86217846935535)) AND (Serina NOT MISSING
AND (Serina > 4.18460680423871))
THEN
        Node = 18
        Prediction = 3
        Probability = 1.000000

/* Node 10 */
IF (Histidina NOT MISSING AND (Histidina > 1.12076082557669 AND Histidina <=
1.39913310456926)) AND (Ácido Glutámico NOT MISSING AND (Ácido Glutámico >
4.86217846935535))
THEN
        Node = 10
        Prediction = 1
        Probability = 0.966667

/* Node 11 */
IF (Histidina NOT MISSING AND (Histidina > 1.39913310456926 AND Histidina <=
1.51311126860383)) AND (Serina NOT MISSING AND (Serina <= 3.6231101511879))
THEN
        Node = 11
        Prediction = 1
        Probability = 0.857143
```

```
/* Node 12 */
IF (Histidina NOT MISSING AND (Histidina > 1.39913310456926 AND Histidina <=
1.51311126860383)) AND (Serina IS MISSING OR (Serina > 3.6231101511879))
THEN
        Node = 12
        Prediction = 3
        Probability = 1.000000

/* Node 4 */
IF (Histidina IS MISSING OR (Histidina > 1.51311126860383))
THEN
        Node = 4
        Prediction = 3
        Probability = 1.000000
```

Anexo 6. Matriz de correlaciones entre los aminoácidos en los Taxa archaeas, bacterias y eucariotes.

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	1.000	-0.199	0.260	0.007	-0.272	0.134	0.072	-0.567	-0.461	-0.134	-0.101	-0.534	0.357	0.367	0.439	-0.374	-0.036	0.152	0.087	-0.524
Sig.		0.000	0.000	0.867	0.000	0.001	0.080	0.000	0.000	0.001	0.013	0.000	0.000	0.000	0.000	0.000	0.375	0.000	0.033	0.000
Cys	-0.199	1.000	-0.320	-0.218	0.075	-0.072	0.368	-0.279	0.105	-0.098	-0.081	0.032	0.324	0.295	-0.037	0.365	0.096	-0.418	0.172	0.240
Sig.	0.000		0.000	0.000	0.068	0.078	0.000	0.000	0.010	0.017	0.048	0.431	0.000	0.000	0.361	0.000	0.019	0.000	0.000	0.000
Asp	0.260	-0.320	1.000	0.782	-0.752	-0.063	-0.439	-0.185	0.319	-0.692	-0.645	-0.012	-0.181	0.215	0.569	-0.600	-0.291	0.637	-0.731	-0.401
Sig.	0.000	0.000		0.000	0.000	0.123	0.000	0.000	0.000	0.000	0.000	0.776	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Glu	0.007	-0.218	0.782	1.000	-0.793	-0.362	-0.566	0.095	0.619	-0.609	-0.646	0.120	-0.346	0.174	0.498	-0.589	-0.478	0.552	-0.881	-0.315
Sig.	0.867	0.000	0.000		0.000	0.000	0.000	0.020	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Phe	-0.272	0.075	-0.752	-0.793	1.000	0.255	0.487	0.255	-0.357	0.574	0.621	0.043	0.002	-0.439	-0.673	0.448	0.257	-0.453	0.772	0.490
Sig.	0.000	0.068	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.296	0.966	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Gly	0.134	-0.072	-0.063	-0.362	0.255	1.000	0.307	-0.179	-0.312	-0.182	-0.001	-0.293	0.272	-0.216	-0.054	-0.008	0.204	0.129	0.281	0.011
Sig.	0.001	0.078	0.123	0.000	0.000		0.000	0.000	0.000	0.000	0.989	0.000	0.000	0.000	0.184	0.852	0.000	0.002	0.000	0.792
His	0.072	0.368	-0.439	-0.566	0.487	0.307	1.000	-0.455	-0.525	0.190	0.228	-0.457	0.591	0.270	0.023	0.250	0.283	-0.425	0.628	-0.031
Sig.	0.080	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.566	0.000	0.000	0.000	0.000	0.442
Ile	-0.567	-0.279	-0.185	0.095	0.255	-0.179	-0.455	1.000	0.506	0.194	0.224	0.607	-0.687	-0.769	-0.627	-0.055	-0.117	0.019	-0.171	0.449
Sig.	0.000	0.000	0.000	0.020	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.179	0.004	0.641	0.000	0.000
Lys	-0.461	0.105	0.319	0.619	-0.357	-0.312	-0.525	0.506	1.000	-0.505	-0.413	0.568	-0.524	-0.227	-0.087	-0.295	-0.359	0.163	-0.654	0.320
Sig.	0.000	0.010	0.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.033	0.000	0.000	0.000	0.000	0.000
Leu	-0.134	-0.098	-0.692	-0.609	0.574	-0.182	0.190	0.194	-0.505	1.000	0.719	-0.074	-0.001	-0.228	-0.431	0.516	0.138	-0.416	0.560	0.044
Sig.	0.001	0.017	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.000	0.071	0.974	0.000	0.000	0.000	0.001	0.000	0.000	0.283
Met	-0.101	-0.081	-0.645	-0.646	0.621	-0.001	0.228	0.224	-0.413	0.719	1.000	0.020	-0.073	-0.349	-0.543	0.381	0.097	-0.375	0.565	0.171
Sig.	0.013	0.048	0.000	0.000	0.000	0.989	0.000	0.000	0.000	0.000		0.626	0.072	0.000	0.000	0.000	0.017	0.000	0.000	0.000
Asn	-0.534	0.032	-0.012	0.120	0.043	-0.293	-0.457	0.607	0.568	-0.074	0.020	1.000	-0.553	-0.376	-0.479	0.131	-0.058	-0.204	-0.284	0.501
Sig.	0.000	0.431	0.776	0.003	0.296	0.000	0.000	0.000	0.000	0.071	0.626		0.000	0.000	0.000	0.001	0.153	0.000	0.000	0.000
Pro	0.357	0.324	-0.181	-0.346	0.002	0.272	0.591	-0.687	-0.524	-0.001	-0.073	-0.553	1.000	0.449	0.314	0.199	0.256	-0.278	0.397	-0.307
Sig.	0.000	0.000	0.000	0.000	0.966	0.000	0.000	0.000	0.000	0.974	0.072	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000
Gln	0.367	0.295	0.215	0.174	-0.439	-0.216	0.270	-0.769	-0.227	-0.228	-0.349	-0.376	0.449	1.000	0.734	-0.069	-0.021	-0.086	-0.104	-0.431
Sig.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.000	0.091	0.606	0.036	0.011	0.000
Arg	0.439	-0.037	0.569	0.498	-0.673	-0.054	0.023	-0.627	-0.087	-0.431	-0.543	-0.479	0.314	0.734	1.000	-0.466	-0.272	0.327	-0.426	-0.604
Sig.	0.000	0.361	0.000	0.000	0.000	0.184	0.566	0.000	0.033	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000
Ser	-0.374	0.365	-0.600	-0.589	0.448	-0.008	0.250	-0.055	-0.295	0.516	0.381	0.131	0.199	-0.069	-0.466	1.000	0.281	-0.521	0.469	0.261
Sig.	0.000	0.000	0.000	0.000	0.000	0.852	0.000	0.179	0.000	0.000	0.000	0.001	0.000	0.091	0.000		0.000	0.000	0.000	0.000
Thr	-0.036	0.096	-0.291	-0.478	0.257	0.204	0.283	-0.117	-0.359	0.138	0.097	-0.058	0.256	-0.021	-0.272	0.281	1.000	-0.242	0.455	0.115
Sig.	0.375	0.019	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.001	0.017	0.153	0.000	0.606	0.000	0.000		0.000	0.000	0.005
Val	0.152	-0.418	0.637	0.552	-0.453	0.129	-0.425	0.019	0.163	-0.416	-0.375	-0.204	-0.278	-0.086	0.327	-0.521	-0.242	1.000	-0.519	-0.335
Sig.	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.641	0.000	0.000	0.000	0.000	0.000	0.036	0.000	0.000	0.000		0.000	0.000
Trp	0.087	0.172	-0.731	-0.881	0.772	0.281	0.628	-0.171	-0.654	0.560	0.565	-0.284	0.397	-0.104	-0.426	0.469	0.455	-0.519	1.000	0.287
Sig.	0.033	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.011	0.000	0.000	0.000	0.000		0.000
Tyr	-0.524	0.240	-0.401	-0.315	0.490	0.011	-0.031	0.449	0.320	0.044	0.171	0.501	-0.307	-0.431	-0.604	0.261	0.115	-0.335	0.287	1.000
Sig.	0.000	0.000	0.000	0.000	0.000	0.792	0.442	0.000	0.000	0.283	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.000	

Anexo 7. Implementación en el Matemática de los calculos necesaris para la partición de las bases de datos en subgrupos y la obtención de los vectores NEC_k.

Base de Vertebrados (no mamíferos)

Datos

- **Generación de los Vectores en la base de vertebrados no mamiferos**

```
Proteo =
  ReadList["C:\\Documents and Settings\\armando\\Escritorio\\Resultados para escribir\\Nuevo
     SPSS con AA\\vertebrados.txt", String, RecordLists -> True];

Proteina = Map[Function[lista , Characters[StringJoin[Characters[Drop[lista, 1]]]]], Proteo];

Length[Proteina]

9387

9387 / 100 // N

93.87

ProteinasParticion = Partition[Proteina, 93];
```

Calculo de las frecuencias de los pares aminoácidos

```
<< Statistics`DataManipulation`

General::obspkg :
  Statistics`DataManipulation` is now obsolete. The legacy version being loaded may conflict with current Mathematica
     functionality. See the Compatibility Guide for updating information. >>
```

```
Combinaciones =
 Map[StringJoin, Flatten[Outer[List, {"A", "C", "D", "E", "F", "G", "H", "I", "K", "L",
     "M", "N", "P", "Q", "R", "S", "T", "V", "W", "Y"}, {"A", "C", "D", "E", "F",
     "G", "H", "I", "K", "L", "M", "N", "P", "Q", "R", "S", "T", "V", "W", "Y"}], 1]]
```

```
{AA, AC, AD, AE, AF, AG, AH, AI, AK, AL, AM, AN, AP, AQ, AR, AS, AT, AV, AW, AY, CA, CC, CD, CE, CF,
 CG, CH, CI, CK, CL, CM, CN, CP, CQ, CR, CS, CT, CV, CW, CY, DA, DC, DD, DE, DF, DG, DH, DI, DK,
 DL, DM, DN, DP, DQ, DR, DS, DT, DV, DW, DY, EA, EC, ED, EE, EF, EG, EH, EI, EK, EL, EM, EN, EP,
 EQ, ER, ES, ET, EV, EW, EY, FA, FC, FD, FE, FF, FG, FH, FI, FK, FL, FM, FN, FP, FQ, FR, FS, FT,
 FV, FW, FY, GA, GC, GD, GE, GF, GG, GH, GI, GK, GL, GM, GN, GP, GQ, GR, GS, GT, GV, GW, GY, HA,
 HC, HD, HE, HF, HG, HH, HI, HK, HL, HM, HN, HP, HQ, HR, HS, HT, HV, HW, HY, IA, IC, ID, IE, IF,
 IG, IH, II, IK, IL, IM, IN, IP, IQ, IR, IS, IT, IV, IW, IY, KA, KC, KD, KE, KF, KG, KH, KI, KK,
 KL, KM, KN, KP, KQ, KR, KS, KT, KV, KW, KY, LA, LC, LD, LE, LF, LG, LH, LI, LK, LL, LM, LN, LP,
 LQ, LR, LS, LT, LV, LW, LY, MA, MC, MD, ME, MF, MG, MH, MI, MK, ML, MM, MN, MP, MQ, MR, MS,
 MT, MV, MW, MY, NA, NC, ND, NE, NF, NG, NH, NI, NK, NL, NM, NN, NP, NQ, NR, NS, NT, NV, NW,
 NY, PA, PC, PD, PE, PF, PG, PH, PI, PK, PL, PM, PN, PP, PQ, PR, PS, PT, PV, PW, PY, QA, QC,
 QD, QE, QF, QG, QH, QI, QK, QL, QM, QN, QP, QQ, QR, QS, QT, QV, QW, QY, RA, RC, RD, RE, RF,
 RG, RH, RI, RK, RL, RM, RN, RP, RQ, RR, RS, RT, RV, RW, RY, SA, SC, SD, SE, SF, SG, SH, SI,
 SK, SL, SM, SN, SP, SQ, SR, SS, ST, SV, SW, SY, TA, TC, TD, TE, TF, TG, TH, TI, TK, TL, TM,
 TN, TP, TQ, TR, TS, TT, TV, TW, TY, VA, VC, VD, VE, VF, VG, VH, VI, VK, VL, VM, VN, VP, VQ,
 VR, VS, VT, VV, VW, VY, WA, WC, WD, WE, WF, WG, WH, WI, WK, WL, WM, WN, WP, WQ, WR, WS, WT,
 WV, WW, WY, YA, YC, YD, YE, YF, YG, YH, YI, YK, YL, YM, YN, YP, YQ, YR, YS, YT, YV, YW, YY}
```

```
FrecParesAA[proteina_] :=
 Partition[CategoryCounts[Map[StringJoin, Partition[proteina, 2, 1]], Combinaciones], 20]
```

```
ParesAAFreqBase = Map[Function[lista, Plus @@ Map[FrecParesAA[#] &, lista]], ProteinasParticion];
```

```
(**AminoAcidFreqBase=Plus@@Map[FreAminoAcid,Proteina]*)
```

```
AminoAcidFreq = Map[Plus @@ # &, ParesAAFreqBase];
```

$$\text{ProbAA} = \frac{\text{AminoAcidFreq}}{\text{Map[Plus @@ \# \&, AminoAcidFreq]}}\ 61\ //\ N$$

```
Export["C:\\Documents and Settings\\armando\\Escritorio\\Resultados
   para escribir\\Nuevo SPSS con AA\\VERTEBRADOS.xls", %]
```

```
C:\Documents and Settings\armando\Escritorio\Resultados
```

Anexo 8. Implementación en el Matemática para la selección aleatoria de las matrices de distancia

Preparación

```
<< Statistics`DataManipulation`
```

```
Combinaciones =
 Map[StringJoin, Flatten[Outer[List, {"A", "C", "D", "E", "F", "G", "H", "I", "K", "L",
     "M", "N", "P", "Q", "R", "S", "T", "V", "W", "Y"}, {"A", "C", "D", "E", "F",
     "G", "H", "I", "K", "L", "M", "N", "P", "Q", "R", "S", "T", "V", "W", "Y"}], 1]]
```

```
{AA, AC, AD, AE, AF, AG, AH, AI, AK, AL, AM, AN, AP, AQ, AR, AS, AT, AV, AW, AY, CA, CC, CD, CE, CF,
 CG, CH, CI, CK, CL, CM, CN, CP, CQ, CR, CS, CT, CV, CW, CY, DA, DC, DD, DE, DF, DG, DH, DI, DK,
 DL, DM, DN, DP, DQ, DR, DS, DT, DV, DW, DY, EA, EC, ED, EE, EF, EG, EH, EI, EK, EL, EM, EN, EP,
 EQ, ER, ES, ET, EV, EW, EY, FA, FC, FD, FE, FF, FG, FH, FI, FK, FL, FM, FN, FP, FQ, FR, FS, FT,
 FV, FW, FY, GA, GC, GD, GE, GF, GG, GH, GI, GK, GL, GM, GN, GP, GQ, GR, GS, GT, GV, GW, GY, HA,
 HC, HD, HE, HF, HG, HH, HI, HK, HL, HM, HN, HP, HQ, HR, HS, HT, HV, HW, HY, IA, IC, ID, IE, IF,
 IG, IH, II, IK, IL, IM, IN, IP, IQ, IR, IS, IT, IV, IW, IY, KA, KC, KD, KE, KF, KG, KH, KI, KK,
 KL, KM, KN, KP, KQ, KR, KS, KT, KV, KW, KY, LA, LC, LD, LE, LF, LG, LH, LI, LK, LL, LM, LN, LP,
 LQ, LR, LS, LT, LV, LW, LY, MA, MC, MD, ME, MF, MG, MH, MI, MK, ML, MM, MN, MP, MQ, MR, MS,
 MT, MV, MW, MY, NA, NC, ND, NE, NF, NG, NH, NI, NK, NL, NM, NN, NP, NQ, NR, NS, NT, NV, NW,
 NY, PA, PC, PD, PE, PF, PG, PH, PI, PK, PL, PM, PN, PP, PQ, PR, PS, PT, PV, PW, PY, QA, QC,
 QD, QE, QF, QG, QH, QI, QK, QL, QM, QN, QP, QQ, QR, QS, QT, QV, QW, QY, RA, RC, RD, RE, RF,
 RG, RH, RI, RK, RL, RM, RN, RP, RQ, RR, RS, RT, RV, RW, RY, SA, SC, SD, SE, SF, SG, SH, SI,
 SK, SL, SM, SN, SP, SQ, SR, SS, ST, SV, SW, SY, TA, TC, TD, TE, TF, TG, TH, TI, TK, TL, TM,
 TN, TP, TQ, TR, TS, TT, TV, TW, TY, VA, VC, VD, VE, VF, VG, VH, VI, VK, VL, VM, VN, VP, VQ,
 VR, VS, VT, VV, VW, VY, WA, WC, WD, WE, WF, WG, WH, WI, WK, WL, WM, WN, WP, WQ, WR, WS, WT,
 WV, WW, WY, YA, YC, YD, YE, YF, YG, YH, YI, YK, YL, YM, YN, YP, YQ, YR, YS, YT, YV, YW, YY}
```

```
FrecParesAA[proteina_] :=
 Partition[CategoryCounts[Map[StringJoin, Partition[proteina, 2, 1]], Combinaciones], 20]
```

```
PosicionesAleatorias[N_] := Module[{Positions = {}, k = 1, Lista},
  While[k ≤ N,
   Lista = RandomInteger[DiscreteUniformDistribution[{1, Length[Proteina]}]];
   Positions = Union[Append[Positions, Lista]];
   k = Length[Positions];
  ];
  Positions]
```

Base de human

Lectura de Ficheros

```
Proteo =
  ReadList["C:\\Documents and Settings\\armando\\Escritorio\\Nueva carpeta\\base proteinas
     FINAL\homo sapiens(824).txt", String, RecordLists -> True];

Proteina = Map[Function[lista , Characters[StringJoin[Characters[Drop[lista, 1]]]]], Proteo];

Length[Proteina]
```

```
824
```

Generación de bases aleatorias

```
positionB1 = PosicionesAleatorias[Round[Length[Proteina] / 2]];
positionB2 = Complement[Range[Length[Proteina]], positionB1];

B1 = Map[Proteina[[#]] &, positionB1];
B2 = Map[Proteina[[#]] &, positionB2];
```

Calculo de las frecuencias de los pares aminoácidos

```
ParesAAFreqBase = Plus @@ Map[FrecParesAA, Proteina]

ParesAAFreqBaseB1 = Plus @@ Map[FrecParesAA, B1];
ParesAAFreqBaseB2 = Plus @@ Map[FrecParesAA, B2];

(**AminoAcidFreqBase=Plus@@Map[FreAminoAcid,Proteina]*)

AminoAcidFreq = Plus @@ ParesAAFreqBase
```

```
{28351, 6469, 17182, 24145, 23224, 31362, 13940, 23442, 19660,
 44289, 13565, 15322, 27112, 15118, 16930, 36344, 29156, 26367, 8646, 14410}
```

```
AminoAcidFreqB1 = Plus @@ ParesAAFreqBaseB1;
AminoAcidFreqB2 = Plus @@ ParesAAFreqBaseB2;

HumanProbAA = AminoAcidFreq / (Plus @@ AminoAcidFreq)
```

$$\left\{\frac{28351}{435034}, \frac{6469}{435034}, \frac{8591}{217517}, \frac{24145}{435034}, \frac{11612}{217517}, \frac{15681}{217517}, \frac{6970}{217517}, \frac{11721}{217517}, \frac{9830}{217517}, \frac{44289}{435034}, \frac{13565}{435034}, \frac{7661}{217517}, \frac{13556}{217517}, \frac{7559}{217517}, \frac{8465}{217517}, \frac{18172}{217517}, \frac{14578}{217517}, \frac{26367}{435034}, \frac{4323}{217517}, \frac{7205}{217517}\right\}$$

```
HumanProbAAB1 = AminoAcidFreqB1 / Plus @@ AminoAcidFreqB1
HumanProbAAB2 = AminoAcidFreqB2 / Plus @@ AminoAcidFreqB2
```

$$\left\{\frac{14545}{222697}, \frac{3501}{222697}, \frac{8586}{222697}, \frac{11871}{222697}, \frac{12030}{222697}, \frac{16068}{222697}, \frac{7308}{222697}, \frac{12009}{222697}, \frac{9716}{222697}, \frac{22651}{222697}, \frac{7060}{222697}, \frac{7852}{222697}, \frac{322}{5179}, \frac{7403}{222697}, \frac{8415}{222697}, \frac{18871}{222697}, \frac{15534}{222697}, \frac{13432}{222697}, \frac{4603}{222697}, \frac{172}{5179}\right\}$$

$$\left\{\frac{1534}{23593}, \frac{2968}{212337}, \frac{8596}{212337}, \frac{12274}{212337}, \frac{11194}{212337}, \frac{5098}{70779}, \frac{6632}{212337}, \frac{3811}{70779}, \frac{9944}{212337}, \frac{21638}{212337}, \frac{6505}{212337}, \frac{830}{23593}, \frac{1474}{23593}, \frac{7715}{212337}, \frac{8515}{212337}, \frac{17473}{212337}, \frac{13622}{212337}, \frac{12935}{212337}, \frac{4043}{212337}, \frac{2338}{70779}\right\}$$

Distancias entre genomas Mitochondriales

Función de distancia

```
VectorB1 = {ArcheaProbAAB1, BacteriasProbAAB1, InvertebradosProbAAB1, InsectosProbAAB1,
   VertebradosProbAAB1, MamíferosProbAAB1, PrimatesProbAAB1, HumanProbAAB1};
VectorB2 = {ArcheaProbAAB2, BacteriasProbAAB2, InvertebradosProbAAB2, InsectosProbAAB2,
   VertebradosProbAAB2, MamíferosProbAAB2, PrimatesProbAAB2, HumanProbAAB2};

HellingerDistance[probA_, probB_] := 4 Plus @@ (Sqrt[probA] - Sqrt[probB])^2
```

- **Ejemplo**

```
Outer[HellingerDistance, {{a, b}, {c, d}}, {{a, b}, {c, d}}, 1]
```

$$\left\{\left\{0, 4\left(\left(\sqrt{a}-\sqrt{c}\right)^2+\left(\sqrt{b}-\sqrt{d}\right)^2\right)\right\}, \left\{4\left(\left(-\sqrt{a}+\sqrt{c}\right)^2+\left(-\sqrt{b}+\sqrt{d}\right)^2\right), 0\right\}\right\}$$

Calculos B1

```
distB1 = Outer[HellingerDistance, VectorB1, VectorB1, 1];

Dimensions[distB1]
```

```
{8, 8}
```

```
distancesB1 = N[Table[distB1[[i, j]], {j, 2, 8}, {i, j - 1}], 20]
```

```
{{0.056769551850081751251}, {0.15294803138580461418, 0.066090265741478410624},
 {0.10619152882382488531, 0.042504079995330121077, 0.015040968493005018159},
 {0.21761042278644496039, 0.10459417195200133832, 0.034200774291243885307,
  0.041622289126128139481}, {0.16862705801567226886, 0.071763953456267650818,
  0.023064116776010084427, 0.020269926303451134035, 0.0089891590482498546036},
 {0.13133754568588274366, 0.043358630863671011325, 0.043176622434387919223,
  0.022172965968055947388, 0.034568825974198251629, 0.013827814957715824776},
 {0.13735157276709162288, 0.056886122895281837508, 0.032577080686008100559,
  0.019261285888708950157, 0.039404723965738107764,
  0.015682568971049613387, 0.012607281416016856208}}
```

```
Export["C:\\Documents and Settings\\armando\\Escritorio\\Matrices de
   distancia con una nueva seleccion aleatoria\Distancias B1-50.xls", distancesB1]
```

```
C:\Documents and Settings\armando\Escritorio\Matrices
  de distancia con una nueva seleccion aleatoria\Distancias B1-50.xls
```

Calculos B2

```
distB2 = Outer[HellingerDistance, VectorB2, VectorB2, 1];
```

```
Dimensions[distB2]
```

```
{8, 8}
```

```
distancesB2 = N[Table[distB2[[i, j]], {j, 2, 8}, {i, j - 1}], 20]
```

```
{{0.050631060221788357451}, {0.14797497871849306194, 0.066596523648894885445},
 {0.10334779935959970888, 0.044331446494170000484, 0.017122419648450457940},
 {0.20326436118149677051, 0.099358578198492921396, 0.030591074824782519989,
  0.037442657828417796130}, {0.16908254933930186694, 0.077108352554242121335,
  0.021905630861192149418, 0.020396767567739404565, 0.0059638887194522693612},
 {0.13560856312957953090, 0.050730671684790134120, 0.035736554664745912323,
  0.020110542678205926607, 0.024526727529355110171, 0.012001541232904994207},
 {0.11471094697844356741, 0.045776611654017124018, 0.032187023269789999829,
  0.015688101996691119855, 0.044998227370084963095,
  0.022230452460699821140, 0.014240994534887630730}}
```

```
Export["C:\\Documents and Settings\\armando\\Escritorio\\Matrices de
   distancia con una nueva seleccion aleatoria\Distancias B2-50.xls", distancesB2]
```

```
C:\Documents and Settings\armando\Escritorio\Matrices
  de distancia con una nueva seleccion aleatoria\Distancias B2-50.xls
```