Agrupación de secuencias genéticas desconocidas en transcriptoma de glándula de veneno de escorpión

Mavelyn Sterling Londoño

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2021

Agrupación de secuencias genéticas desconocidas en transcriptoma de glándula de veneno de escorpión

Mavelyn Sterling Londoño

Trabajo de grado presentado como requisito parcial para optar al título de:
Estadística

Director:
Santiago Castaño, PhD
Codirector:
Jaime Mosquera, PhD
2021

Resumen

El análisis de secuencias genéticas es muy útil para conocer más sobre su composición, como identificar en las secuencias de ADN, los aminoácidos compuestos por combinaciones de 4 bases nitrogenadas que luego se traducen y conforman las proteínas. Este trabajo muestra la importancia del estudio de la similaridad entre las secuencias genéticas para identificar patrones que permitan conocer su composición y agruparlas. Dado que la base de datos que se utilizó son secuencias genéticas desconocidas provenientes del transcriptoma de la glándula de veneno de escorpión, primero se realizó un análisis bibliométrico para identificar el crecimiento de los estudios en esta área y cuales son los trabajos más influyentes en esta investigación. Posteriormente, se implementaron técnicas multivariantes como el análisis de componentes principales (ACP) y análisis de conglomerados para caracterizar las secuencias y agruparlas de acuerdo a la similaridad en su composición.

Palabras clave: Similaridad de secuencias genéticas, Medidas de distancia, *k*-tuplas, agrupación de secuencias.

Abstract

The analysis of genetic sequences is very useful to know more about their composition, such as identifying in DNA sequences, amino acids composed of combinations of 4 nitrogenous bases that are then translated and form proteins. This work shows the importance of studying the similarity between genetic sequences to identify patterns that allow us to know their composition and group them. Since the database used are unknown genetic sequences from the scorpion venom gland transcriptome, first a bibliometric analysis was performed to identify the growth of studies in this area and which are the most influential works in this research. Subsequently, multivariate techniques such as principal component analysis (PCA) and cluster analysis were implemented to characterize the sequences and group them according to similarity in their composition.

Keywords: Similarity of genetic sequences, Distance measures, *k*-tuples, sequence grouping

Contenido

Lista de Figuras

Lista de Tablas

Declaración

Afirmo que he realizado el presente Trabajo de Grado de manera autónoma y con la única ayuda de los medios permitidos y no diferentes a los mencionados en el propio trabajo. Todos los pasajes que se han tomado de manera textual o figurativa de textos publicados y no publicados, los hemos reconocido. Ninguna parte del presente trabajo se ha empleado en ningún otro tipo de Tesis o Trabajo de Grado.

Igualmente declaro que los datos utilizados en este trabajo están protegidos por las correspondientes cláusulas de confidencialidad.

Santiago de Cali, 08.10.2021

_____

(Mavelyn Sterling Londoño)

Introducción

Una secuencia genética es una sucesión de combinaciones de 4 bases nitrogenadas (T, C, G, A) que define la información genética de las células. Existen dos tipos de secuencias genéticas: el ADN (ácido desoxirribonucleico), que contiene las instrucciones para la producción de proteínas y el ARN (ácido ribonucleico), que permite a la célula leer y comprender las instrucciones del ADN. Estas cadenas de texto se estudian con el fin de encontrar patrones e identificar similitudes con secuencias conocidas de estudios anteriores, con lo cual se obtiene mayor información sobre la funcionalidad de las secuencias genéticas en las proteínas.

La agrupación y la identificación de las similitudes entre las secuencias genéticas es muy importante para la identificación de patrones en secuencias poco estudiadas. Este es el caso de las secuencias provenientes de la glándula del escorpión, el cual es uno de los animales más venenosos, por lo que la composición de su veneno es objetivo de estudio para la construcción del perfil genético y posibles antisueros y vacunas.

Las dos glándulas secretoras del veneno se encuentran en una vesícula llamada Telson, las cuales están cubiertas de una membrana y separadas por un tabique muscular vertical. La comunicación de la glándula de veneno con el exterior, es por medio de pliegues que limita el ingreso de luz pero que se abre minutos antes de la picadura con el fin de expulsar el veneno (Polis 1990).

Para el análisis de las secuencias genéticas del transcriptoma de la glándula veneno, se tienen más de 60.000 secuencias desconocidas, es decir, que no han presentado alguna similitud con las secuencias que ya se tienen registro en la mayor base de datos biológica *GenBank*. Estas secuencias pueden contener información valiosa para estudios biológicos y genéticos posteriores, que pueden servir para la producción de vacunas y medicamentos.

De acuerdo con lo anterior, el propósito de este trabajo es identificar similitudes en las secuencias genéticas desconocidas de la glándula del veneno de escorpión. Para ello se utilizarán herramientas estadísticas tales como el análisis multivariado de componentes principales, que permiten resumir la información de múltiples variables en unas pocas componentes, en donde se quiere estudiar la relación de las variables con las secuencias genéticas, y ayuda a identificar patrones que facilitan posibles agrupamientos, por lo que se combina con otra técnica que es la de conglomerados.

En esta labor, cada secuencia es caracterizada a través de variables tales como la longitud total de la secuencia, la distribución de frecuencia relativa de los nucleótidos y la distribución de frecuencias relativas de las combinaciones de 2 y 3 de los nucleótidos. Luego de conformar los grupos, se estudia cada grupo particular identificando las características de similaridad al interior de cada grupo y las diferencias frente a las secuencias entre los grupos.

El presente documento se encuentra construido de la siguiente manera: en el capítulo 1 está la definición del problema, justificación y objetivos; el capitulo 2 se presenta la revisión de antecedentes con la búsqueda bibliográfica y el análisis bibliométrico; el capitulo 3 se describe el marco teórico compuesto del marco conceptual biológico y el marco teórico estadístico; el capitulo 4 muestra la metodología empleada para el análisis de las secuencias genéticas; el capítulo 5 se presentan los resultados obtenidos; el capítulo 6 se encuentran las conclusiones y las recomendaciones para investigaciones futuras.

1 Definición del problema

1.1. Planteamiento del problema

El ADN (ácido desoxirribonucleico) contiene un conjunto de instrucciones necesarias para la producción de las proteínas que determinan el funcionamiento de las células. Un paso intermedio entre la secuencia de ADN y una proteína determinada, es la transcripción, que se define como el proceso en el cual se cambia el sistema de codificación de ADN al de ARN. El análisis de las secuencias de ARN (ácido ribonucleico) se le denomina transcriptoma (Dawkins 1982).

En los análisis de los transcriptomas se comparan las secuencias de ARN (Adenina, Uracilo, Citosina, Guanina) obtenidas de una muestra frente a un archivo histórico de secuencias almacenadas y caracterizadas en bases de datos internacionales. Por ejemplo, *"National Center for Biotechnology Information"*, *"European Nucleotide Archive"*, entre otras, son de las bases de datos biológicas internacionales más grandes. Estas bases de datos, contienen información de las secuencias genéticas (de ADN, ARN o proteínas) que han sido reportadas, tales como las secuencias provenientes del genoma de organismos, veneno de animales, virus, entre otros. Luego de realizar la comparación frente a las bases de datos, las secuencias son etiquetadas como conocidas o desconocidas dependiendo de su similaridad. Las secuencias conocidas son utilizadas para crear el perfil de la expresión genética, mientras que las secuencias desconocidas suelen ser excluidas de los estudios, lo que implica una posible pérdida de valiosa información, que puede ser relevante en desarrollos científicos posteriores (Schwartz et al. 2007).

Las bases de datos de secuencias ARN se encuentran almacenadas en un formato FASTA, utilizado para almacenar grandes volúmenes de cadenas de texto. Cada secuencia tiene una longitud y composición o combinación distinta de su código (A,U,C,G). Desde la bioinformática se han utilizado técnicas y metodologías consideradas convencionales tanto para el manejo de base de datos como para la detección de patrones en las secuencias (Escobar et al. 2011).

Un ejemplo de estudios genéticos, es relacionado con el análisis de las secuencias genéticas provenientes de la glándula de veneno del escorpión. Estas investigaciones son relativamente recientes (Schwartz et al. (2007), Ma et al. (2009) y Alvarenga et al. (2012)), por lo cual,

hay poca información sobre la composición de estas secuencias. Sin embargo, la necesidad de conocer y explorar las secuencias desconocidas es cada vez mayor para desarrollos científicos como vacunas, medicamentos, entre otros. Actualmente existen aproximadamente 60000 secuencias genéticas de la glándula del veneno de escorpión que son desconocidas y conocidas al rededor de 2600 aproximadamente. Por lo que, el análisis de todas las secuencias de ADN desconocidas es muy importante para poder encontrar similitudes entre ellas o con las secuencias ya reportadas en las bases de datos internacionales, lo que permite agrupar las secuencias con características similares por su composición de nucleótidos.

1.2. Justificación

De acuerdo con la Organización Mundial de la Salud (OMS), el envenenamiento producido por picaduras de escorpión es un problema de salud pública, debido que se han identificado algunas especies con veneno mortal para los seres humanos como *Hadrusrus gertschi* (Schwartz et al. 2007), *Tityus serrulatus* (Alvarenga et al. 2012), *Hemiscorpius lepturus* (Kazemi-Lomedasht et al. 2017). La escasez de antisueros incrementa el número de muertes, por lo que sugirió a nivel internacional, en especial en países tropicales, hacer un análisis de estructuras proteicas de las toxinas con el propósito de producir antivenenos (OMS 2007).

En el área de las ciencias ómicas se han realizado estudios sobre la glándula de veneno de algunas especies de escorpiones como lo muestran Schwartz et al. (2007) y Ma et al. (2009), las cuales fueron las primeras investigaciones utilizando escorpiones. Estos trabajos lograron identificar moléculas y péptidos (cadenas de aminoácidos) que han orientado la elaboración de antibióticos y antitumorales que pueden ser útiles, por ejemplo para infecciones por *Escherichia coli* y para el cáncer de próstata y colón (Villamil 2016).

El análisis de transcriptomas proporciona información sobre los procesos biológicos que ocurren en las células de la glándula del veneno de escorpión. La identificación de las secuencias conocidas y desconocidas de ARN contribuyen a estudios genéticos posteriores, ya que se podrían identificar más patrones de expresión. Estos patrones permiten conocer más información sobre las componentes del veneno y así, crear un perfil genético para la elaboración de antibióticos naturales (Alvarenga et al. 2012).

En la literatura nacional, las publicaciones de investigaciones de análisis estadísticos con bases de datos biológicas son escasos, por lo que se hace necesario ampliar el conocimiento de las tendencias de investigación en esta área, por medio de un análisis bibliométrico, con el fin de obtener información sobre las investigaciones con mayor impacto en esta área, los autores más relevantes, el volumen de artículos y citaciones de las publicaciones, para seleccionar adecuadamente los antecedentes de este trabajo.

Las secuencias genéticas son vectores de combinaciones de 4 letras (bases nitrogenadas), las cuales pueden ser objeto de análisis de estadístico, ya que se puede obtener para cada secuencia, las distribuciones de frecuencias de la aparición de las bases nitrogenadas y su respectiva longitud, siendo estas un conjunto de variables. En consecuencia, cada registro de la base de datos de secuencia desconocida tendrá asociado las variables que las caracterizan tales como las distribuciones de frecuencia de las bases nitrogenadas, las cuales pueden ser analizadas de forma conjunta a través de métodos de análisis multivariado, tales como el análisis de componentes principales, y generar agrupación a través métodos de análisis de conglomerados.

1.3. Objetivos

1.3.1. Objetivo general

Identificar y caracterizar los transcriptos de secuencias no conocidas presentes en glándula de veneno de escorpión, mediante modelos estadísticos multivariados.

1.3.2. Objetivos específicos

- Identificar las tendencias y el crecimiento de la literatura sobre el análisis de secuencias genéticas.

- Realizar un análisis exploratorio de las secuencias genéticas desconocidas encontradas en un transcriptoma de glándula de veneno de escorpión.

- Agrupar el conjunto de secuencias desconocidas en subgrupos de secuencias estadísticamente similares.

- Identificar y caracterizar los patrones especiales en los grupos de secuencias desconocidas de la glándula del veneno de escorpión.

2 Revisión de antecedentes

En este trabajo la revisión de los antecedentes ha sido planeada para desarrollarse en dos etapas. En la primera etapa se realiza una búsqueda bibliográfica extensa, cuyas características generales son analizados a través de un análisis bibliométrico apoyado en la aplicación *Biblioshiny* de la librería de Bibliometrix del software R (Aria & Cuccurullo (2017)). El objetivo del análisis es identificar las tendencias y la producción científica sobre el análisis de secuencias genéticas. Superada esta fase, aquellos trabajos que se identifican como más relevantes, son estudiados en su contenidos en una revisión bibliográfica más detallada.

2.1. Búsqueda Bibliográfica y Análisis Bibliométrico

Se realizó una búsqueda bibliográfica extensa sobre los trabajos más relevantes en dos bases de datos bibliográficas (*SCOPUS* y *Web Of Science - WOS*) relacionados con el análisis de secuencias genéticas, específicamente se consideran trabajos relacionados con la evaluación de la similaridad y la agrupación de secuencias. La búsqueda se limitó a revistas de las áreas de *ciencias de la computación, ingeniería, matemática aplicada, biología, ciencia de la decisión y multidisciplinaria*, bajo la siguiente ecuación de búsqueda:

ALL((“Sequence similarity”) AND (“Distance measure” OR “alignment-free comparison”)) AND (LIMIT-TO (SUBJAREA , “COMP”) OR LIMIT-TO (SUBJAREA , “BIOC”) OR LIMIT-TO (SUBJAREA , “MATH”) OR LIMIT-TO (SUBJAREA , “ENGI”) OR LIMIT-TO (SUBJAREA , “DECI”) OR LIMIT-TO (SUBJAREA , “MULT”))

En ambas bases de datos bibliográficas se tomaron los trabajos más citados entre los años 2005 y enero del 2021. En *SCOPUS*, se obtuvo una base de datos con 278 documentos, mientras que en *WOS* se identificaron solo 18 artículos. Posteriormente, utilizando la función *biblioshiny* del paquete Bibliometrix -software R- Aria & Cuccurullo (2017), se obtuvieron los indicadores de relevancia para la producción científica de ambas fuentes. Estos indicadores son:

- Promedio de citas por año.

- Autores más relevantes.
- Número total de apariciones de los autores.
- Número de artículo por autor.
- Número de autores por artículo.
- Índice de colaboración.

Para realizar el análisis de las redes de colaboración, se unieron las dos hojas de datos, eliminando los registros repetidos. Sobre esta hoja de datos consolidados, se realizó el análisis de los siguientes indicadores.

- Los países con más artículos citados.
- Las revistas científicas con más artículos.
- Índice H.
- Red de colaboración entre autores.
- Red de co-ocurrencia de palabras claves de los autores.

Algunos documentos claves, encontrados en esta búsqueda bibliográfica, fueron estudiados con mayor profundidad e incluidos en el marco teórico y de referencia.

Descripción	**SCOPUS**	**WOS**
Total de artículos	278	18
Fuentes (revistas, libros, etc.)	178	11
Promedio de citas por artículo	24	25.5
Número de autores	762	69
Apariciones de los autores	955	71
Artículos por autor	0.365	0.261
Autores por artículo	2.74	3.83
Co-Autores por artículos	3.44	3.94
Índice de colaboración	2.86	3.83

Tabla 2-1: Información principal de las fuentes bibliográficas *SCOPUS* y *Web Of Science* sobre análisis de secuencias genéticas.

La Tabla **2-1**, se resume la información general sobre los trabajos encontrados en la búsqueda bibliográfica. Estos trabajos tuvieron un promedio de 24 citas por artículo en *SCOPUS* y 25.5 en *Web Of Science*. Se destaca que el número de autores por trabajo es

relativamente alto, 2.74 en *SCOPUS* y 3.83 en *WOS* lo que a su vez se refleja en altos índices de colaboración 2.86 y 3.83 respectivamente.

Para evaluar la producción científica de los autores, se unieron las dos bases de datos provenientes de SCOPUS y WOS. Los registros duplicados fueron identificados a través del campo DOI y posteriormente eliminados. Luego de realizar este filtro, el número de trabajos se redujo a 275.

La Figura **2-1** representa el crecimiento en la producción científica a partir del año 2005 hasta el año 2021. Según esta figura, es notable el crecimiento desde el 2005 al 2014, siendo el periodo 2014-2015 de mayor número de publicaciones (30 y 28 artículos por año). A pesar de este crecimiento global, es importante resaltar que en el transcurso de los años la producción en esta área se ha sostenido, es decir que no hay variaciones muy grandes en la cantidad de artículos.

Debido que, en la base de datos se hizo corte de documentos publicados hasta enero del año 2021, solo se tiene un artículo.

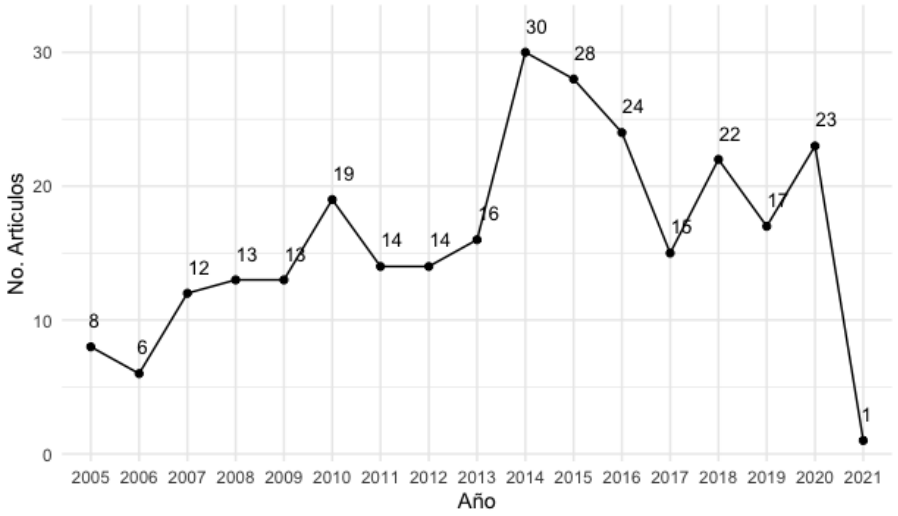

Figura 2-1: Producción científica anual sobre análisis de secuencias genéticas

2.1.1. Análisis por volumen de citaciones

Se seleccionaron los 10 países con más citas en la base de datos (Figura **2-2**). Estados Unidos (USA) es uno de los países más influyentes, con un total de 1485 citas, seguido de Italia con 1086 citas, mientras Polonia, con 142 citas, es el país de menor influencia de este grupo.

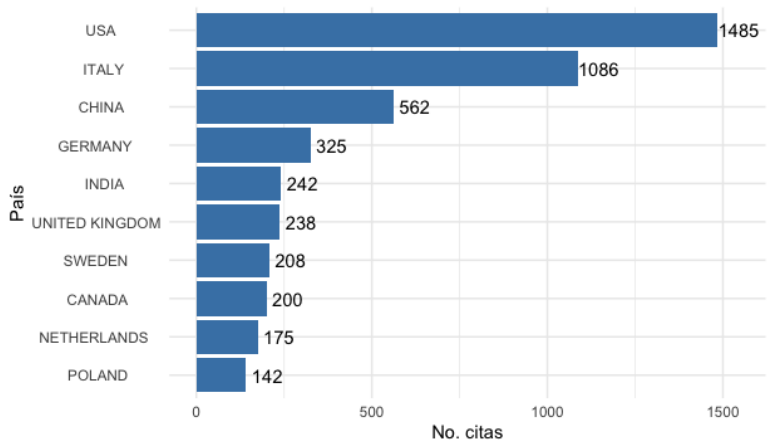

Figura 2-2: Los 10 países con mayor volumen de producción científica sobre análisis de secuencias genéticas.

En la Figura **2-3**, se presenta los 10 autores con mayor producción científica y una representación de la dinámica de las citaciones obtenidas a través en el periodo 2006-2021. El autor *Matteo Comin*, perteneciente a la Universidad de Padova-Italia, es quien tiene más publicaciones ya que ha producido 12 documentos. Sus publicaciones empezaron en el año 2014, y el 2019 fue el año en que tuvo el promedio de citas más alto. *András Kocsor*, de la Universidad de Szeged, y *Sandor Pongor*, de la Universidad Católica Pázmány Péter en Hungría, se destacaron en el 2008 por sus publicaciones y el promedio de citas, al igual que *Alberto Apostolico* y *Cinzia Pizzi* en el año 2016, que fueron los que realizaron más publicaciones de artículos.

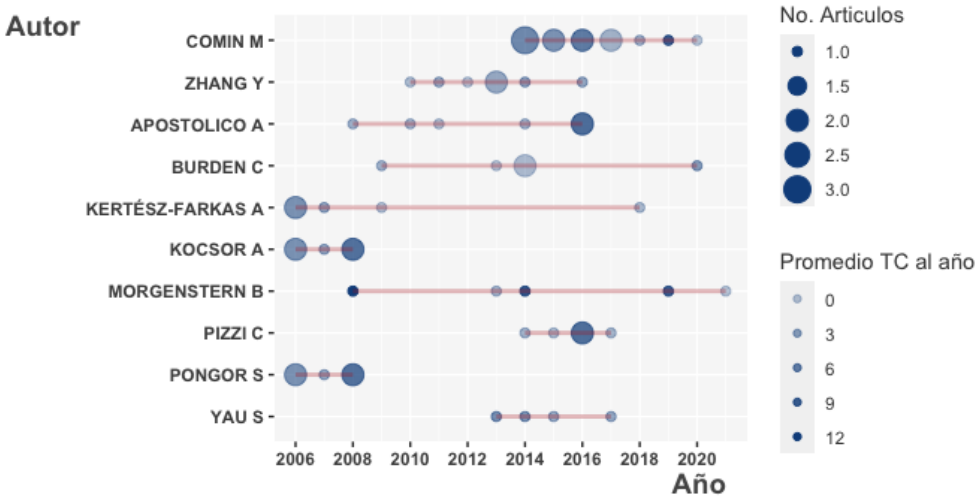

Figura 2-3: Producción científica anual de los 10 autores más citados.

2.1.2. Análisis del impacto de las colaboraciones

Para medir el impacto de la producción científica, se presentan las 10 revistas con mayor cantidad de artículos. De acuerdo con la Figura **2-4**, la revista *Bioinformatics* tiene un total de 15 artículos publicados sobre esta temática, seguido de *Lecture Notes in Compute Science* con 10 artículos, mientras *IEEE Transactions on multimedia* y *Leibniz Internacional Proceeding in Informatics Lipics* presentan 4 documentos cada uno, siendo las revistas menos influyentes en este conjunto de revistas científicas.

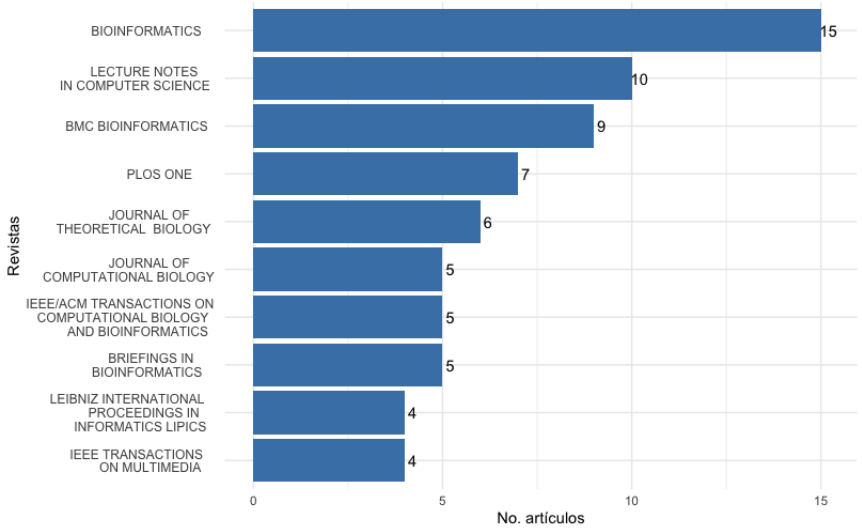

Figura 2-4: Las 10 revistas científicas con más publicaciones sobre análisis de secuencias genéticas.

En la Tabla **2-2**, se presenta un indicador bibliométrico (índice H) para los 10 autores con mayor producción científica (ver la Figura **2-3**), acompañados del total de citas de los artículos publicados, el total de trabajos publicados y el año en cual realizó la primera publicación; *Matteo Comin* presenta el mayor índice H = 7: Esto quiere decir que 7 artículos de los 12 trabajos publicados han sido citados al menos 7 veces; el autor *Burkhard Morgenstern*, tiene el mayor número de citas con un total de 285, en 5 artículos publicados.

Autor	Índice H	Total de Citas	Total de artículos	Primer año
COMIN M	7	144	12	2014
ZHANG Y	6	86	7	2010
APOSTOLICO A	5	99	6	2008
BURDEN C	3	31	5	2009
KERTÉSZ-FARKAS A	3	109	5	2006
KOCSOR A	5	204	5	2006
MORGENSTERN B	4	285	5	2008
PIZZI C	4	66	5	2014
PONGOR S	5	204	5	2006
YAU S	4	102	4	2013

Tabla 2-2: Indicadores bibliométricos de los 10 autores más relevantes sobre análisis de secuencias genéticas.

2.1.3. Redes de colaboración

La red de colaboración de los autores permite identificar la colaboración entre los investigadores o autores en la producción científica. Esta red se obtiene matemáticamente por la teoría grafos, en donde la red o el grafo se compone de nodos o vértices que son los nombres de los autores, y los enlaces o aristas son las co-autorías, por lo que se obtiene una red bipartita, es decir, una red simple, que se puede dividir en dos grupos distintos compuestos por vértices, pero los vértices de un mismo conjunto no pueden estar relacionados, resultando una matriz de dimensión: *documentos* x *autores* (Aria & Cuccurullo 2017).

De acuerdo con la Figura **2-5**, se observa que los autores *Matteo Comin* y *Burkhard Morgenstern*, son los de mayores producción científica. Sin embargo, *Matteo Comin* no ha realizado colaboraciones con otros autores influyentes, mientras que *Burkhard Morgenstern* y *Fengzhu Sun* presentan una alta relación de coautoría.

La red de Co-ocurrencia al igual que red de colaboración, se construye por la teoría de grafos, pero ahora los vértices son las palabras claves de los artículos estudiados. En la Figura **2-6**, muestra las 22 palabras claves más frecuentes. Particularmente, se observa que las palabras "*Clustering*", *"Sequence Comparison"* y "*Alignment-free"* son las más utilizadas. Por lo cual, se puede verificar que el proceso de la revisión bibliográfica fue acorde con al cual se desea buscar y analizar los trabajos realizados.

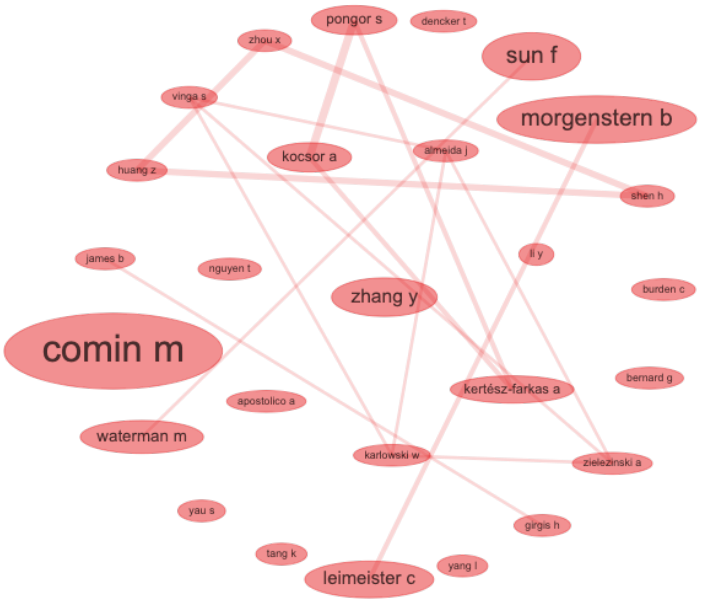

Figura 2-5: Red de colaboración entre los autores más influyentes sobre análisis de secuencias genéticas.

2.1.4. Selección de trabajos relacionados con el análisis de secuencias genéticas

Por medio del análisis realizado anteriormente, se encontraron algunos artículos de investigación, que plantean, implementan y validan metodologías sobre el análisis de secuencias genéticas. Como es el caso de las combinaciones de las *k-tuplas* (subcadenas de longitud k) de letras (A, T, C, G), los cuales permiten establecer de manera formal y matemática el algoritmo y modelo a seguir para el desarrollo de este trabajo. La revisión a detalle de estos artículos se presenta en la sección de 2.2.

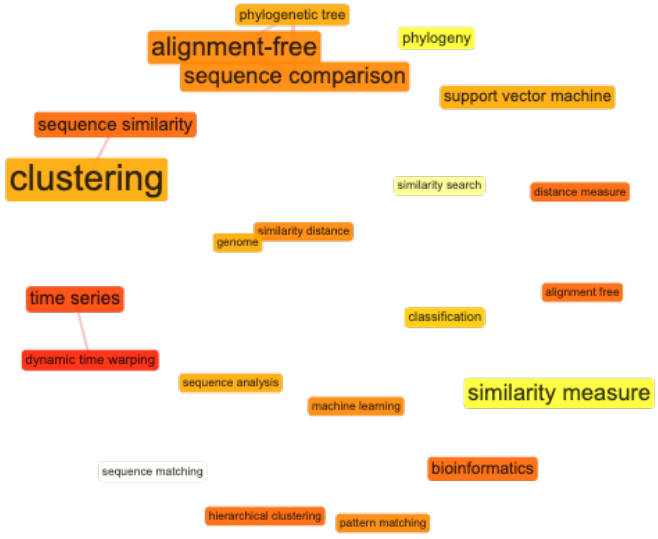

Figura 2-6: Red de co-ocurrencia de palabras claves de las publicaciones sobre análisis de secuencias genéticas.

2.2. Revisión Bibliográfica

La presente sección se orienta a identificar y describir los trabajos que en el análisis bibliométrico previo (sección 2.1), han realizado la tarea de agrupar secuencias genéticas. Posteriormente, la revisión hace especial énfasis en los trabajos previos que abordaron el problema de agrupación de secuencias del transcriptoma de la glándula del veneno de escorpión. A pesar de que hay varios estudios sobre las secuencias del transcriptoma de la glándula de veneno de organismos venenosos, pocos son sobre el veneno de escorpión, ocasionando que la información existente sobre las secuencias que componen este veneno sea limitada.

Vinga & Almeida (2003), presentan una revisión de dos metodologías para comparar secuencias genéticas. El primer método se basa en la frecuencia de palabras en una secuencia, para lo cual analizan la distribución de frecuencias de las k-tuplas (k= *1, 2, 3, 4*) que se presentan en toda la secuencia de ADN (A = Adenina, T = Timina, C = Citosina, G = Guanina). Bajo este enfoque, la métrica de distancia entre dos secuencias corresponde a la distancia euclidiana y a la entropía relativa de las distribuciones de las frecuencias de las k-tuplas que se conforman en las dos secuencias. Estas métricas mostraron ser útiles para comparar secuencias de gran longitud, ya que se dividen en subcadenas cadenas de texto. En contraste, el segundo método no requiere dividir la secuencia completa en fragmentos de secuencias que se utilizan para el proceso de transcripción para la producción de proteínas, sino que se basa en la representación de secuencias mediante mapas iterativos y de la

complejidad de Kolmogorov. Estas métricas fueron propuestas por Almeida et al. (2001) y Li et al. (2001), pero su desempeño solo ha sido estudiado sobre casos de ejemplos limitados; el primer enfoque de frecuencias de palabras se puede utilizar para comparar secuencias de ADN, proteínas y textos en lenguaje natural. El segundo enfoque, solo lo recomiendan para comparar proteínas, sugiriendo estudios posteriores para ampliar la aplicación el algoritmos a más tipos de secuencias genéticas.

Baena et al. (2006), realizaron una revisión de las técnicas de clustering empleadas para la clasificación de expresiones genéticas, formando grupos de datos con similitudes en su composición o función celular, por lo que esta clasificación se considera como no supervisada, debido a que los grupos no están predefinidos. Las secuencias genéticas que utilizaron fueron almacenadas en microarrays, el cual es un pequeño chip que permite coleccionar miles de genes respecto a una serie de condiciones experimentales, obteniendo como resultado una matriz donde las filas corresponden a los genes y las columnas a las condiciones experimentales y las celdas son marcadas por un color que muestra el grado de expresión genética en cada condición experimental. Posteriormente, utilizaron técnicas de clustering convencionales como *K-means, SOM y técnicas jerárquicas como el método de grupo de pares no ponderados con media aritmética (UPGMA)*, que son útiles cuando se conoce el número de grupos a obtener, y no convencionales como *Identificación de clústeres a través de núcleos de conectividad* (CLICK), *Técnica de búsqueda de afinidad de clústeres* (CAST) y *un método de agrupamiento jerárquico basado en densidad* (DHC), las cuales mostraron tener un mejor resultado en problemas con número de grupos desconocido. Además, hacen una comparación entre las técnicas de agrupación con los algoritmos de Biclustering, estableciendo que estos algoritmos son óptimos cuando se tiene cuenta con una base de datos pequeña de genes, ya que este algoritmo tiene como objetivo identificar grupos de secuencias de genes que tengan patrones similares en su funcionamiento o actividad celular con un subconjunto de restricciones experimentales.

Comin et al. (2015), realizaron una revisión de las métricas para medir la distancia entre las distribuciones de probabilidad de frecuencias de las k-tuplas conformadas en dos secuencias genéticas. Una de estas métricas, llamada *"D_2"*, emplea la correlación entre el número de ocurrencia de todas las k-tuplas en dos secuencias, es decir el producto interno entre dos vectores de palabras (X e Y), cada uno representando el número de veces que aparecen las palabras de las k-tuplas en las dos secuencias. Sin embargo, se menciona que esta métrica puede presentar sesgo por el ruido estocástico de cada secuencia, por lo cual, el conjunto de métricas de distancias llamadas D son variaciones de D_2. El desempeño de estas métricas fue evaluado sobre conjuntos de secuencias de ARN humano extraídas de NCBI (*National Center for Biotechnology Information*), seleccionando al azar 50 conjuntos con 100 secuencias cada uno. Además, midieron la calidad de la lectura de las secuencias por medio de una máquinas de secuenciación, donde los puntajes que se muestran como la

probabilidad de que la lectura de cada secuencia sea correcta.

Posteriormente, por medio del paquete llamado *"QCluster"* desarrollado en C^{++}, se agruparon las secuencias con el algoritmo k-means, junto con la norma de la distancia euclidiana, divergencia de Kullback-Liebler y D_2.

Amiri & Dinov (2017), Implementaron y validaron métodos para análisis de secuencias genéticas, por medio de la librería *msktuple* en el lenguaje de programación R, la cual representa y calcula la distancia entre las secuencias. Entre las métricas empleadas para medir la distancia entre secuencias genéticas sin alineación se encuentran *locational k-tuple*, el cual calcula la distancia para cada letra, teniendo en cuenta las posiciones en las que aparece en la secuencia, *naive k-tuple* que se basa en la distancia Euclídea y *CV-Tree* que utiliza la correlación entre dos secuencias. El desempeño de estas métricas fue probado con secuencias genéticas de 9 mamíferos distintos extraídas del repositorio *GenBank*. Aunque obtuvieron resultados satisfactorios, los autores recomiendan seguir explorando en técnicas de análisis de secuencias para poder elegir la metodología más adecuada al problema de estudio.

Konishi et al. (2019), agruparon un conjunto de secuencias genéticas utilizando dos enfoques. En el primero, realizaron una conversión numérica para calcular la matriz de distancia entre dos secuencias utilizando la distancia euclidiana en variables de dos letras juntas. Posteriormente, realizaron un ACP (*Análisis de componentes principales*), ya está técnica no asume una estructura en los datos, y lo que hace es encontrar direcciones de diferencias, las cuales se representan en ejes. Posteriormente, aplicaron una de las técnicas cluster (Cluster jerárquico y no jerárquico). Para el segundo enfoque, las letras de las secuencias se transforman a lenguaje booleano (0 y 1), donde 1 representa la presencia de cierta letra y 0 la ausencia de la letra obteniendo una matriz booleana, la cual se rota utilizando la descomposición de valores singulares, para encontrar las componentes principales e implementar un algoritmo de agrupación.

Análisis de transcriptoma de la glándula de veneno de escorpión

Schwartz et al. (2007), realizaron un análisis de las transcripciones de la glándula de veneno del escorpión mexicano *Hadrusrus gertschi*, el cual no se considera peligroso para los humanos y únicamente se encuentra en lugares como túneles excavados en el estado de Guerrero, México. Para esto, se utilizó una estrategia para obtener la secuenciación de genes, por medio de una biblioteca de secuencias ADN_c (ADN complementario), la cual es construida a partir del ARN extraído de algún tejido o glándula. Anteriormente se habían estudiado los perfiles de las transcripciones de algunos organismos venenosos, como lo son los cnidarios, serpientes, arañas y caracoles cónicos. Este es el primer estudio en que se

analiza la glándula de veneno de escorpión.

Para construir la biblioteca de ADN_c (ADN complementario) se extrajeron las secuencias de ARN del telson, ubicado en la parte baja del abdomen del escorpión y que contiene las glándulas del veneno; estas secuencias genéticas son clonadas, para así generar 160 etiquetas de las secuencias cortas expresadas EST (*Expressed Sequence Tags*). Luego se emplearon algoritmos de búsqueda bioinformáticos (Blastn y Blastx), comparando con las 147 EST que resultaron de alta calidad, agrupando las secuencias por la similitud en su composición, las cuales son importantes para la construcción del perfil genético.

Después de la búsqueda, se identificaron que un total de 68 grupos de secuencias únicas y conocidas, con los cuales se logró una descripción de las componentes del veneno y se amplio el conocimiento del perfil del veneno para estudios posteriores. Mientras 43 grupos de secuencias no lograron ser identificadas, las cuales fueron eliminadas en el procedimiento de identificar las componentes del veneno.

Ma et al. (2009), caracterizaron las secuencias genéticas conocidas del transcriptoma de la glándula de veneno del escorpión *Jendeki*, perteneciente de la provincia de Yunnan, ubicado en el suroeste de China. Estos autores plantearon realizar el trabajo con un enfoque en secuencias cortas expresadas EST (*Expressed Sequence Tags*), con la finalidad de obtener un panorama general del transcriptoma de la glándula venenosa, de una muestra de 50 escorpiones *Jendeki*. Para esto utilizaron un método para el análisis de las secuencias de ARN, en el cual se construye una biblioteca de ADN_c extraído del ARN de la glándula de veneno. Estas secuencias de ADN_c se clonan para obtener las secuencias EST, de las cuales resultaron 871 secuencias. Además se establece un perfil de las expresión génica, es decir, reconocer las componentes del veneno del escorpión.

Posteriormente, se empleó un análisis cluster utilizando la distancia euclidiana entre dos secuencias genéticas para clasificar las secuencias genéticas obtenidas, comparándolas con las secuencias reportadas en las bases de datos *SWISS-PROT* y *GenBank (NCBI)*. Además recurriendo al algoritmo Blast, el cual realiza búsquedas por similitud entre una secuencia obtenida y todas las bases de datos disponibles, teniendo en cuenta parámetros específicos como la longitud y la composición de las secuencias. Como resultado, se obtuvieron 208 grupos con secuencias ya conocidas, de los cuales 33 grupos son secuencias que presentan transcripciones similares a los péptidos del veneno del escorpión *Jendeki*. Mientras 175 grupos no son similares, y 85 grupos no se lograron establecer alguna coincidencia con las secuencias reportadas del escorpión *Jendeki*, las cuales fueron descartadas y podrían tener información valiosa del veneno.

En este trabajo se lograron identificar 10 tipos de toxinas conocidas y 9 tipos atípicos,

obteniendo información importante sobre el veneno de escorpión, que podría ayudar en el desarrollo de medicamentos y vacunas.

Alvarenga et al. (2012), realizaron un estudio transcriptómico de la glándula de veneno del escorpión *Tityus serrulatus*, proveniente de Brasil. Para esto construyeron una biblioteca de secuencias ADN_c a partir de las secuencias de ARN extraídas de la glándula de veneno de 60 escorpiones, las cuales son clonadas y cultivadas por 18 horas en medio selectivo con antibióticos. Posteriormente se secuencian y etiquetan, obteniendo las secuencias EST *(expressed sequence tags)* con la ayuda del equipo para analizar secuencias genéticas llamado *ABI 3130*. Para mejorar la calidad de las secuencias EST clonadas, se utilizó el programa *SeqClean*, el cual elimina las secuencias denominadas adaptadores o enlazadoras con otras secuencias, obteniendo las secuencias con mayor puntuación en la transcripción y así realizar una clasificación por grupos empleando un análisis cluster.

Posteriormente las secuencias se registran en la base de datos *GenBank* para compararlas con las secuencias de la base de datos *UnitProt*, utilizando un algoritmo bioinformático de búsqueda (BLASTx). Permitiendo encontrar que 1259 secuencias EST son proteínas conocidas, con las cuales, se creó el perfil de la transcripción, identificando tres secuencias completas de toxinas y varias proteínas de la glándula de veneno de escorpión, que contribuyen a la elaboración de vacunas de mordeduras con veneno. En contraste, 370 secuencias EST no tuvieron alguna coincidencia, por lo que fueron descartas pero pueden ser parte de proteínas nuevas que no se han identificado.

En conclusión, el análisis bibliométrico y la revisión bibliográfica ha permitido encontrar y estudiar la metodología más adecuada para el análisis de secuencias genéticas, por medio de la distribución de frecuencias de las k-tuplas. Además, identificar los trabajos que han realizado en los últimos años sobre la agrupación de secuencias genéticas y la importancia de la aplicación de las técnicas multivariadas para estos análisis. Permitiendo establecer la metodología más adecuada para desarrollar en este trabajo, estimando la distribución de las *k-tuplas*, realizando un análisis de componentes principales y luego una agrupación de las secuencias de acuerdo a su similaridad entre las secuencias.

3 Marco Teórico

En este capitulo se presentan de forma resumida los términos y conceptos más importantes requeridos para comprender el desarrollo de la metodología y los resultados obtenidos en el presente proyecto. Este marco teórico es dividido en dos partes. La primera parte es un marco conceptual, que muestra las definiciones alusivas a la terminología celular, análisis de transcriptoma, bioinformática y la relación de la estadística en los análisis de datos biológicos; en la segunda parte se presenta el *marco teórico estadístico*, en el cual se describen las métricas de distancias entre secuencias genéticas, el análisis de componentes principales y las técnicas de agrupación multivariadas.

3.1. Marco conceptual

3.1.1. La célula

La teoría celular fue propuesta en 1839 por el fisiólogo Theodor Schwann y el botánico Matthias Jakob Schleiden. Esta teoría hizo un gran impacto en la biología moderna, planteando que todos los seres vivos contiene una o más células, que constituyen la unidad funcional más pequeña de los seres vivos, en las cuales se almacena toda la información genética. Tiempo después, en 1855, el médico y biólogo alemán Rudolf Virchow planteó la tercera idea de esta teoría, que plantea que las células transmiten su material genético a nuevas células a través de la división celular, llevando a cabo el proceso de reproducción celular (Starr & Taggart 2008).

Existen dos tipos de células, debido a su tamaño, forma y constitución, que son las *procariotas* y *eucariotas*. Las células procariotas carecen de un núcleo y su estructura es simple, por lo que su material genético se encuentra en el *nucleoide*. Este tipo de células están solamente presentes en las bacterias. Los otros organismos como hongos, animales y plantas poseen células eucariotas, las cuales tienen una estructura compleja, ya que contiene un *núcleo* encargado de recolectar el material genético y variedad de orgánulos necesarios para llevar a cabo los procesos metabólicos. Sin embargo, ambas clases de células presentan semejanzas en el funcionamiento (Cooper & Hausman 2010).

Las células se pueden ver como un sistema de información, donde se almacena y se procesa un conjunto de datos provenientes de una misma fuente. En el caso de la célula eucariota, la fuente es el núcleo en el que se encuentra almacenado el ADN (ácido desoxirribonucleico) que son secuencias genéticas. Su estructura fue descubierta por Francis Crick y James Watson en 1953, las cuales tienen información que permiten determinar y llevar a cabo su funcionamiento. Estas secuencias son sistemas que funcionan por combinaciones de 4 bases nitrogenadas, para el caso del ADN son: *Adenina, Timina, Citosina y Guanina*. Su representación es por pares de bases, es decir una doble hélice, donde la *Adenina* se conecta con la *Timina*, y la *Guanina* con la *Citosina*. En cambio, para el ARN (ácido ribonucleico) sus bases nitrogenadas son *Adenina, Guanina, Citosina y Uracilo*, por lo que la *Adenina* ahora se conecta con el *Uracilo* (Karp 2005).

La representación del ADN es distinta a la del ARN, ya que es una cadena compuesta en su interior por las cuatro bases (A,U,C,G) y una capa en la superficie de azúcar fosfato, por lo que la longitud de las secuencias de bases nitrogenadas determina la información genética almacenada (Cooper & Hausman 2010).

Para la producción de las proteínas, que van a componer a su vez los orgánulos y estas a la misma célula, se requiere de unas instrucciones que están almacenadas en el ADN, como si fuese un sistema de computo parecido al lenguaje binario (0 y 1) de las computadoras. Para ejecutar estas instrucciones lo primero que se debe hacer es transcribir cada instrucción y ese transcrito es llamado *ARNmensajero (ARNm)*, el cual se debe "perfeccionar" para salir del núcleo hacia el retículo endoplasmático en donde los ribosomas ayudan a pasar esas instrucciones a proteínas. Este proceso se le denomina *Traducción* (Karp 2005).

En el proceso de la transcripción hay regiones de las secuencias del ADN que no son codificadas para la producción de proteínas, llamadas *intrones*, y las que son codificadas se llaman *exones*. El término anteriormente utilizado como "perfeccionar", hace referencia al proceso de eliminar los *intrones* de forma tal que solo se queden secuencias de *exones*, obteniendo un *ARNm maduro*, ya que dependiendo de los tipos de *exones* y sus posibles combinaciones será el tipo de proteína que se obtiene (Karp 2005).

De acuerdo con los estudios genéticos realizados por Gregor Mendel en 1865, sobre el estudio de herencia, se puede entender un *gen* como un conjunto de instrucciones necesarias en la célula con el fin de generar una función o característica específica; estos se reconocen porque tienen un patrón de inicio con la secuencia de las bases AUG (*Adenina*, *Uracilo* y *Guanina*) y otro patrón de finalización dado por una secuencia de 4 o más *Adeninas* juntas, a estos patrones se les llama péptido señal y lo que se encuentra en medio de ellos se conoce como marco abierto de lectura, es decir la instrucción. Sin embargo, pueden haber instrucciones que no sigan esta estructura o que sus péptidos señales todavía no sean reconocidos, y que

estos contribuyan a explicar procesos que aún no se han logrado entender por completo (Karp 2005).

3.1.2. Bioinformática

Desde finales de siglo XX la Bioinformática ha tenido un gran avance y revolución en el campo científico, ya que reúne las áreas biológicas e informática; esta disciplina utiliza algoritmos computaciones y manejo de bases de datos para analizar y caracterizar las proteínas y las secuencias genéticas que constituyen a un organismo (Genoma). Para esta tarea se desarrollan herramientas de programación que se relacionan con problemas biológicos como enfermedades, evolución, expresiones genéticas, entre otros (Escobar et al. 2011).

Dentro de la Bioinformática se han identificado 3 escuelas o enfoques distintos; la primera escuela es la *formal*, en la cual las prácticas de simulación y búsqueda son enfocadas en la inteligencia y vida artificial, ampliando las líneas de investigación a temas como algoritmos genéticos, redes neuronales, teoría de autómatas celulares, entre otros. La segunda es la escuela *estructural*, que se encuentra enfocada en los estudios de secuencias de proteínas, ácidos nucleicos (ADN y ARN) llamada *ciencias ómicas*, que tienen como finalidad identificar genes de una secuencia de ADN, agrupar secuencias genéticas con patrones similares, y analizar las relaciones evolutivas por medio de árboles filogenéticos obtenidos de proteínas. Por último, esta la escuela *instrumental*, que se encarga del manejo y gestión de bases de datos biológicos ubicadas en redes locales llamadas *LANs*, utilizando búsquedas de información como documentos de literatura científica por medio de similitud semántica (Lahoz-Beltrá 2010).

A partir de las 3 escuelas mencionadas, se puede medir el alcance de la aplicación de técnicas computacionales en proyectos para el desarrollo de medicamentos, vacunas, análisis de secuencias genéticas, entre otros.

3.2. Marco teórico estadístico

Los estudios en la bioinformática y estadística han tenido un alto incremento, obteniendo más información y mejores algoritmos útiles para la comunidad científica como es en el caso de los estudios genéticos. Para empezar, una secuencia genética es una combinación de 4 bases nitrogenadas, las cuales son almacenadas en el formato *FASTA* (Fast-A, A de una secuencia de ADN) que permite almacenar grandes cadenas de texto. En este formato, la primera linea corresponde a la procedencia de las secuencias como el nombre del animal

y le sigue cadenas de texto que son las grandes secuencias (Figura **3-1**) (Escobar et al. 2011).

Existen variables o características que se pueden medir sobre las secuencias genéticas, como su longitud, la frecuencia de ocurrencia de cada una de las bases nitrogenadas y de las combinaciones de 2 o más bases. Cuando la secuencia genética es agrupada en subconjuntos de k-caracteres estos reciben el nombre de k-tuplas. Por lo cual, el análisis de la distribución de frecuencia de las k-tuplas permite caracterizar cada secuencia genética y evaluar la similaridad entre pares de secuencias para agruparlas.

```
>TRINITY_DN36600_c0_g1_i1 len=226 path=[0:0-225]
CCTTTGGCAGTTATTATTATTACAGCAGAAATCTGTGTGGGCATTATCTTGATTAACTTCTGTAACATTGCCACTTTTATATTGTACATT
TCTAATTCAGAACCCTGGCTAGCTTCTGGGGGCTGTTGGGATGAGTTGTTCTACTTCTGAGTTGGAAATTCAATTCATCCCAGAGGTGTT
CAATGGGATTAAGGTCTGGACTCTGGGCAGACTAGTTGAATCTTGC
```

Figura 3-1: Ejemplo de una secuencia genética en el formato FASTA.

3.2.1. k-tuplas en secuencias genéticas

Para caracterizar una secuencia genética se espera que las 4 bases nitrogenadas no se distribuyan aleatoriamente, sino que sigan algún patrón identificable. Para estudiar estos patrones se especifica el número de caracteres a agrupar (k-tuplas) para encontrar alguna de similaridad entre las secuencia genéticas (Figura **3-2**).

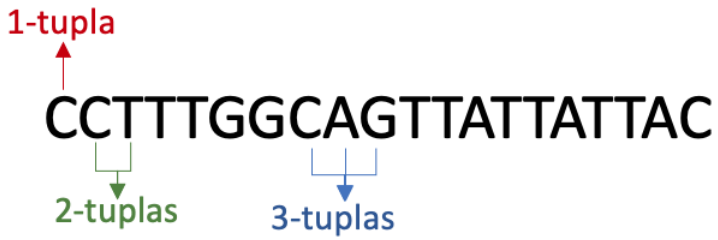

Figura 3-2: Ejemplo de las posibles subcadenas de longitud k de una secuencia genética.

Análisis de distribución de frecuencias

En el análisis de secuencias genéticas por las k-tuplas, se considera que cada secuencia tiene una longitud ($\mathbb{L}$), desde la cual se obtiene la frecuencia relativa de cada posible combinación de la k base (k-tuplas) de las secuencias. Es decir, se mide la cantidad de veces en la que se repite en la secuencia de k letras. En este sentido, la cantidad de variables que se tendrían de acuerdo a las k-tuplas se resume a 4^k. Por ejemplo, la distribución de frecuencia para k=1 tendrá un rango de variación de tamaño 4, para k= 2 el rango de variación es de tamaño 16

(Figura **3-3**) y para k=3 se pueden conformar 64 posibles combinaciones de 3 letras. En la práctica, el orden máximo de análisis de la k-tuplas suele ser k=3.

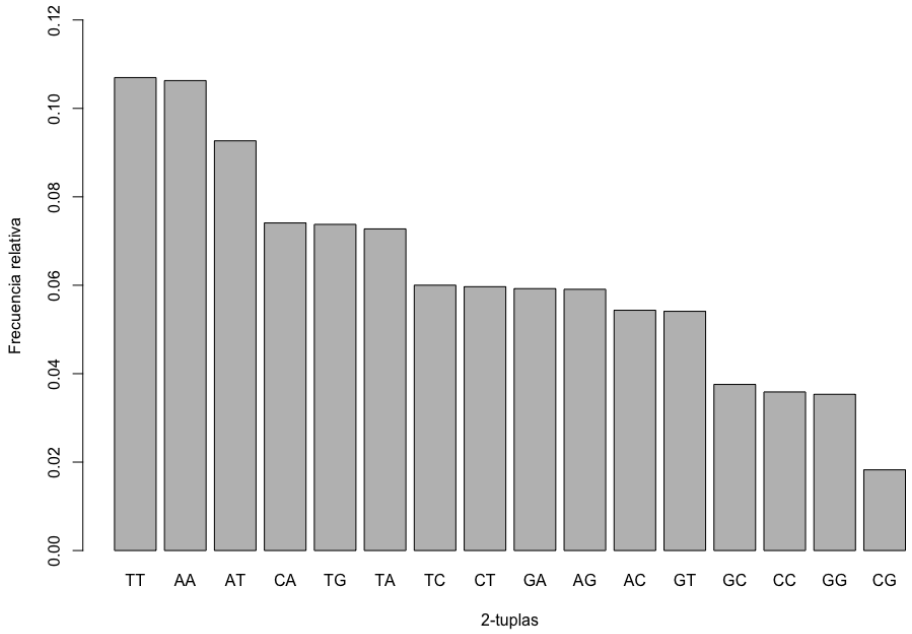

Figura 3-3: Ejemplo de la distribución de frecuencia de 2-tuplas de una secuencia genética.

Medidas de distancia entre objetos

En este tipo de análisis los datos están representados por una medida de proximidad, es decir, de la distancia o métrica entre 2 elementos de la base de datos. Una métrica o distancia es un función entre dos puntos x y $y \in \mathbb{R}^d$, a la que se le asocia un valor positivo y entre más alto sea, más distante o alejados se encuentran los puntos o elementos, por ejemplo la distancia euclídea entre las k-tuplas (Rodriguez & Proyecto 2015).

Distancia entre las k-tuplas

Cuando se tienen p=2^k variables, que son la distribución de frecuencia relativa de las k-tuplas de dos secuencias genéticas, cada una con diferente longitud (Amiri & Dinov 2017). La similaridad se puede calcular en función de la distancia Euclídea, que es la raíz cuadrada de la suma de las diferencias de las distribuciones de frecuencia de las k-tuplas elevada al cuadrado.

$$D(S_1, S_2) = \sqrt{\sum_{i=1}^{p} (S_{1i} - S_{2i})^2}$$

Siendo S_{1i} y S_{2i} las frecuencias relativas de la i-ésima combinación del conjunto de las secuencias S_1 y S_2 respectivamente.

Disimilitud a través de CV-Tree

A diferencia de la distancia euclídea, esta medida se basa en la correlación entre las secuencias pareadas (Amiri & Dinov 2017).

$$\rho(S_1, S_2) = \frac{\sum_{i=1}^{p} S_{1i} S_{2i}}{\sqrt{\sum_{i=1}^{p} (S_{1i})^2 \sum_{i=1}^{p} (S_{2i})^2}}$$

Esta correlación puede también ser insumo para calcular otra medida de disimilitud entre dos secuencias genéticas (Amiri & Dinov 2017).

$$D(S_1, S_2) = \frac{1 - \rho(S_1, S_2)}{2}$$

3.2.2. Análisis de Componentes Principales (ACP)

El ACP fue planteado inicialmente por Karl Pearson en 1901, sin embargo en 1933, Harold Hotelling retomó el trabajo de Pearson integrando la estadística matemática, el cual planteó Análisis de Componentes Principales con el fin de estudiar la relación entre las variables y los individuos al mismo tiempo (Lebart et al. 1995).

La técnica del ACP es la más importante del análisis multivariado, la cual se utiliza para representar los datos de manera óptima de acuerdo a criterios geométricos y matemáticos, para interpretar y comprender las relaciones que existen entre las variables y los individuos en una matriz de datos X_{nxp}, con n número de observaciones y q número de variables.

El ACP permite reducir la dimensión, es decir, transformar la información contenida en una matriz de datos con p variables originales a un conjunto menor q variables incorrelacionadas (q < p), llamadas Componentes Principales (φ).

$$\begin{aligned}
\varphi_1 &= u_{11}X_1 + u_{21}X_2 + ... + u_{p1}X_p \\
\varphi_2 &= u_{12}X_1 + u_{22}X_2 + ... + u_{p2}X_p \\
&\vdots \\
\varphi_q &= u_{1q}X_1 + u_{2q}X_2 + ... + u_{pq}X_p
\end{aligned}$$

Las q componentes principales se obtienen por combinaciones lineales con las variables originales, y se ordenan según el porcentaje de varianza explicada de cada componente. Por lo cual, la primera componente siempre es la más importante, ya que explica el mayor porcentaje de variabilidad de las observaciones. Donde λ_j son valores propios (j = 1,..., q) de la matriz de covarianzas o correlaciones $\sum$ de orden p. En el ACP se puede utilizar dos tipos de matrices: la matriz de correlaciones, en casos donde todas las variables se encuentran en diferente escala de medida y tienen la misma importancia, y la matriz de covarianza, la cual se emplea cuando las variables tienen la misma escala de medida y se quiere estudiar las variables de acuerdo al grado de variabilidad que puedan presentar (Ochoa Muñoz 2018).

$$\lambda_1 > \lambda_2 > \lambda_3 > ... > \lambda_q$$

En ACP es muy importante poder maximizar la correlación que existe entre los nuevos ejes con las variables originales, y maximizar la variabilidad o la inercia que se obtiene en cada eje α. La inercia se calcula con la siguiente fórmula:

$$\varphi'_\alpha M_n^{-1} \varphi_\alpha = u' Z' M_n Z u = \lambda_\alpha$$

Teniendo en cuenta la restricción de que $u'u = 1$, y que $M_n = 1/n$ es la matriz diagonal ponderada a cada individuo. Por lo cual, la matriz que se va a diagonalizar es $Z' M_n Z = uDu'$, donde D es la matriz diagonal que contiene los valores propios, u son los vectores propios y Z es la estandarización de la matriz de datos X.

Los valores y vectores propios surgen cuando se maximiza la expresión de la inercia utilizando el Lagrangiano $L(u)$, con el propósito que los valores propios sean decrecientes y garantizar que las componentes principales φ sean ortogonales.

$$\begin{gathered} L = u' Z' M_n Z u - \lambda(u'u - 1) \\ L = u' Z' M_n Z u - \lambda u'u + \lambda \\ \frac{\partial L}{\partial u} = 2 Z' M_n Z u - 2\lambda u = 0 \\ \frac{\partial L}{\partial u} = 2(Z' M_n Z u - \lambda u) = 0 \\ Z' M_n Z u = \lambda u \end{gathered}$$

Donde la matriz $Z' M_n Z u = cor(X_i, X_j)$ es simétrica.

Para la interpretación en el ACP se observa la relación entre las componentes y las variables originales, por medio del plano cartesiano, estudiando el signo y la magnitud de las correlaciones de las variables (Figura **3-4**-a). Estas correlaciones al cuadrado son llamadas contribución relativa, que mide la variabilidad de las variables sobre cada eje o dimensión. En cambio, los individuos se representan por medio de nubes de puntos en las dimensiones del espacio $\mathbb{R}^n$, en el cual, el centro de gravedad es el origen (Figura **3-4**-b).

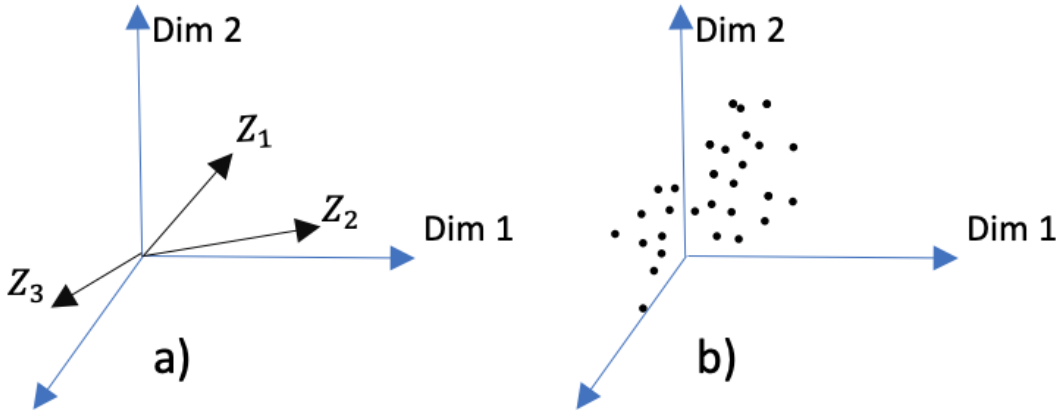

Figura 3-4: Representación de las variables y de los individuos en un plano de las componentes principales.

Análisis de la nube de individuos

Una forma de interpretar de manera gráfica a los individuos de una base de datos, es proyectándolos en un plano factorial para observar mejor la nube de individuos. Sin embargo, para lograr analizar la relación de los individuos con las Componentes Principales, se debe implementar una estrategia, una puede ser encontrar un subespacio de una dimensión H que logre maximizarlas proyecciones de la suma de cuadrados de las distancias de todos los pares de individuos $(i,\, i^{'})$ (Lebart et al. 1995).

$$Max_H \left\{ \sum_{i}^{n} \sum_{i'}^{n} d^2(i, i') \right\}$$

Siendo equivalente a maximizar la suma de cuadrados de la distancia entre el individuos i con el centro de gravedad G del plano (Lebart et al. 1995).

$$Max_H \left\{ \sum_{i}^{n} d^2(i, G) \right\}$$

Una de las distancias más utilizadas es la euclidiana. Dado que en las bases de datos puede ocurrir que las variables no se encuentran en la misma escala de medición, es importante hacer una corrección a las escalas, por lo se estandariza las variables j (Lebart et al. 1995).

$$d^2(i,i') = \sum_{j=1}^{p} \left(\frac{x_{ij} - x'_{ij}}{s_j \sqrt{n}} \right)^2$$

Donde s_j es la desviación estándar de la variable j.

$$s_j^2 = \frac{1}{n} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{ij})^2$$

Uno de los indicadores para la interpretación del ACP, es la contribución de los individuos en un eje de inercia o variabilidad, es decir que se mide la importancia de el individuo u observación en el eje α.

$$CTR(i,\alpha) = \frac{\varphi_{i,\alpha}^2}{n\lambda_\alpha}$$

$$\sum_{i=1}^{n} CTR(i,\alpha) = 1$$

Análisis de la nube de variables

En el caso de las variables, su análisis se hace respecto al punto de origen O del plano, es decir que la distancia de las variables y el origen es igual a 1 (Lebart et al. 1995).

$$d^2(j,O) = \sum_{i=1}^{n} (x_{ij}^2) = 1$$

Ahora bien, los puntos de las variables se encuentran sobre un circulo con radio 1, llamado *el circulo de correlaciones*. Las variables que mejor se representan son las más cercanas al circulo de correlaciones. Respecto a las ejes que representan las componentes principales, las variables que más se asocian a cada eje son lo que presentan un mayor valor como se observa en la Figura **3-6** (Lebart et al. 1995).

En el espacio R^n, el coseno de un ángulo θ entre los vectores de dos variables es el coeficiente de correlación y dado que la distancia con el origen es 1, los cosenos del ángulo son su producto escalar, es decir que si dos variables se encuentran a una misma dirección es porque presentan una correlación positiva, mientras que si dos variables se encuentra en dirección contraria su correlación es negativa (Lebart et al. 1995).

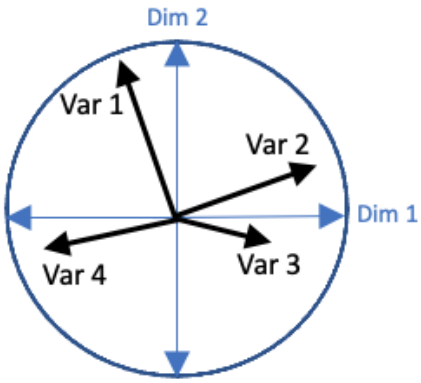

Figura 3-5: Ejemplo de las variables sobre el circulo de correlaciones en ACP.

Otra ayuda para la interpretación de las variables es la contribución de una variable j al eje α, el cual tiene como objetivo, detectar las variables que más ayudan a la construcción de los ejes factoriales y la suma de las contribuciones en un eje α es igual a 100. Estas contribuciones se calculan teniendo en cuenta la proyección de los puntos-fila de la variable j (φ_j), que se obtienen en la relación de transición, en cual se conoce la direcciones factoriales de un espacio y calcular las direcciones factoriales en otro espacio (Lebart et al. 1995).

$$CTR(j,\alpha) = \frac{\varphi_{j\alpha}^2}{\lambda_j} = u_{j,\alpha}^2$$

$$\sum_{j=1}^{p} CTR(j,\alpha) = 100$$

Los cosenos cuadrados (Cos^2) también permiten interpretar la nube de variables en el plano factorial, ya que se puede conocer la calidad de la representación, por lo cual, entre más cerca esté una variable del círculo de correlaciones, mejor será su representación en el plano factorial (Kassambara 2017). En el caso de que una variable se encuentre casi perfectamente representada por solo dos componentes principales (Dim 1 y Dim 2), la suma de Cos^2 en estas dos es igual o cercano a 1.

$$CTR(j,\alpha) = \frac{\varphi_{j\alpha}^2}{Var_j}$$

$$\sum_{\alpha=1}^{p} Cos^2(j,\alpha) = 1$$

Representación simultánea

Evaluar la relación de la nube de individuos con las variables es por medio de la representación en el plano de gráficos biplots (Figura **3-6**). En el cual, se identifican las variables que más caracterizan en los individuos, así como las que menos representan a los individuos, permitiendo conocer más el comportamiento de los individuos respecto a las variables y posteriormente agrupar los individuos con más cercanos.

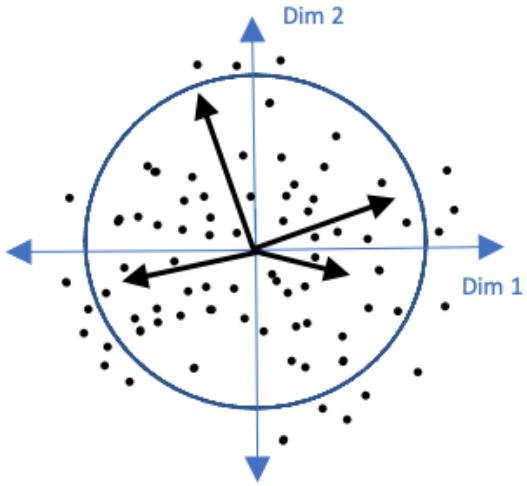

Figura 3-6: Representación simultanea de las variables y los individuos en un plano factorial.

3.2.3. Análisis conglomerados

El análisis de conglomerados, o también llamado análisis cluster, se encarga de dividir un conjunto de datos u observaciones multivariadas en subconjuntos que se denominan *clases* o *grupos (cluster)*. En cada grupo o clase sus elementos contienen características similares que se diferencian de los otros grupos, por lo que hay diferentes técnicas de agrupación que buscan obtener información sobre patrones en los datos; estas técnicas son empleadas en diversas áreas como en el procesamiento de imágenes, búsquedas en la web, marketing, estudios biológicos, estadística, aprendizaje automático, entre otros (Han et al. 2011).

Tipos de técnicas de conglomerados

- **Agrupación jerárquica:** Este algoritmo agrupa datos con característica similares, pero las clases o grupos se conforman en un orden determinado llamado *Jerarquía*, el cual se divide en dos métodos. El primero es el *divisivo* donde inicialmente todas

las observaciones pertenecen a un solo grupo, luego se va dividiendo en más grupos con similitudes, y esto se repite de forma recursiva en cada grupo o cluster, hasta que cada observación tenga su propio grupo. El segundo es el *aglomerativo*, este método es contrario al anterior, ya que primero se le asigna un cluster a cada observación, luego se calcula la similitud o distancia entre cada par de grupos, y los dos grupos que presenten menos distancia se fusionan obteniendo un nuevo grupo, este proceso se repite hasta lograr que todas los datos se encuentren en un solo grupo. El algoritmo se puede consultar en (Han et al. 2011, pág. 460, capítulo 10, figura 10.6).

La representación gráfica de una agrupación jerárquica entre observaciones o individuos es por medio de un árbol llamado dendrograma, el cual puede estar de forma horizontal o vertical. Las hojas del dendrograma hace referencia a los individuo u observaciones, y a medida que se asciende en el árbol, las ramas se van fusionando de acuerdo a su similaridad conformando grupos se puede observar en la Figura **3-7**. Los grupos que se conformaron más rápido son aquellos en los que sus observaciones son más similares entre ellos, mientras los grupos que se forman más lento tienden a ser diferentes. Para determinar el número de grupos óptimos en un dendrograma se debe realizar un corte, que se realiza a la altura de cortes de las fusiones más grandes.

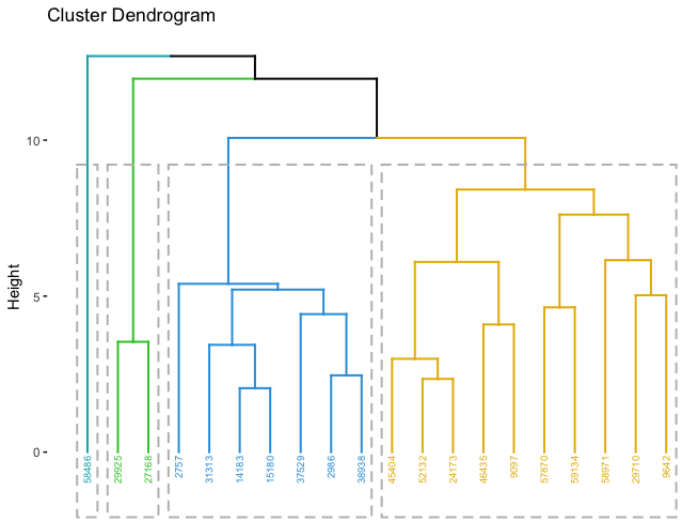

Figura 3-7: Ejemplo de un dendrograma con 20 observaciones.

- **K-means:** Fue propuesto por Steinhaus en 1957. Este método es el más utilizado en el análisis de conglomerados con variables de tipo cuantitativo, y está basado en el principio de mínimos cuadrados y en el modelo de regresión de mínimos cuadrados. Su algoritmo es iterativo, por lo que consta de un ciclo en el cual se selecciona i

observaciones centrales llamadas *centroides*, y son definidos como estado inicial. Luego a cada observación del total de los datos, se calcula la distancia elegida y a la más cercana (el valor más bajo) se le asigna el grupo, después se actualiza los grupo donde se requiere encontrar un nuevo centroide, pero ahora promediando las observaciones que pertenecen a cada grupo. Este proceso se repite hasta encontrar el centroide después de clasificar más datos sea el mismo (Figura **3-8**). El algoritmo se puede consultar en (Han et al. 2011, pág. 452, capítulo 10, figura 10.2).

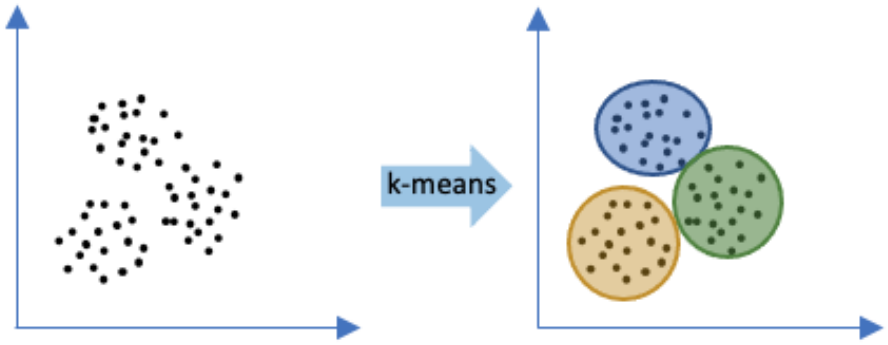

Figura 3-8: Agrupación por el algoritmo k-means.

Número óptimo de grupos

Para determinar el número adecuado de grupos cuando no se tiene información apriori como referencia, se puede utilizar el método del codo (elbow method) que consiste en aplicar el algoritmo de k-means a un rango de valores de K grupos e identificar a partir de cual valor K la suma total de varianza internas de los grupos parece estabilizarse, ya que este valor es el número adecuado de de grupos (Han et al. 2011).

El algoritmo de k-means, considera que los mejores grupos son quienes tienen una la variabilidad interna más baja, ya que es un problema de optimización en el cual al agruparse los individuos en los K grupos, la suma de las varianzas internas tiene que ser lo menor posible. La varianza interna de un grupo K ($W(C_K)$), se calcula como la suma de las distancias euclídeas al cuadrado entre cada individuo u observación (x_i) y el centroide del grupo (μ_K) (Han et al. 2011).

$$W(C_K) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

Dado que los algoritmos de clasificación, como el k-means, son algoritmos iterativos, cada vez que se ejecute puede arrojar una agrupación distinta, por lo que es necesario estabilizar este proceso para obtener la mejor agrupación posible. Para esto, es importante tener en cuenta un criterio de convergencia, como puede ser el de la varianza de intra-clases que decrece hasta permanecer constante, entre la etapa (m) en que se determinaron nuevos centroides, teniendo en cuenta, las primeras observaciones centrales como centro de gravedad, y la etapa siguiente (m+1) , en la que otros nuevos centroides.

Caracterización de los grupos

Al tener las observaciones o individuos agrupados, se calcula las distribuciones de frecuencia relativa de 1-tupla, 2-tuplas y 3-tuplas de cada grupo, para observar las tuplas con mayor frecuencia. Luego, esas distribuciones de frecuencia son relacionadas con el diagrama de aminoácidos, encontrando los aminoácidos que más comunes en cada grupo.

4 Metodología

En este capítulo se resume la metodología planteada para realizar el análisis y la agrupación de las secuencias genéticas desconocidas en transcriptoma de glándula de veneno de escorpión. Dado lo novedoso del tema de estudio, inicialmente se realizó una búsqueda bibliográfica extensa, cuyo resultado se analizó a través de las herramientas del análisis bibliométrico. Esta primera fase se desarrolló con el apoyo de los buscadores *SCOPUS*, *Web Of Science (WOS)* y las herramientas de análisis disponibles en la librería Bibliometrix del software R. Los resultados de este análisis bibliométrico fueron descritos en el Capítulo 2 del presente documento.

Posteriormente se realizó la de depuración de las secuencias genéticas, en la cual se estudió la estructura de las secuencias para construir la distribución de frecuencias de la letras según su aparición en la secuencia y la distribución de la frecuencia relativa para 1-tupla, 2-tuplas y 3-tuplas. Posteriormente, se realizó un análisis estadístico, el cual está conformado por el análisis multivariante y técnicas cluster del aprendizaje no supervisado. En la Figura **4-1**, se ilustra el diseño metodológico que se realizó por fases.

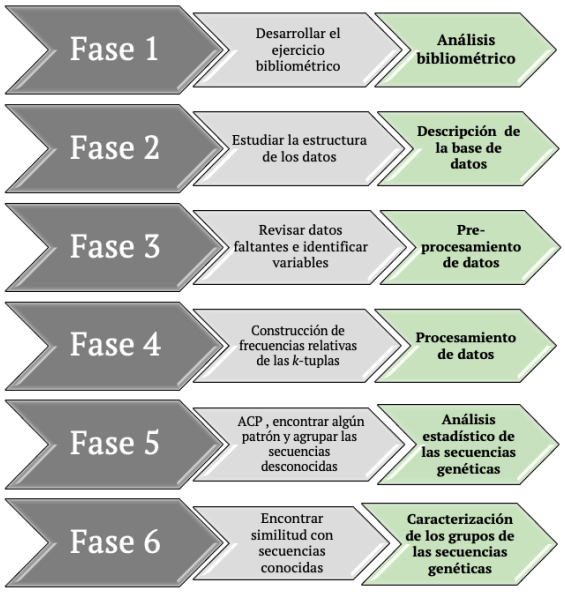

Figura 4-1: Diagrama por fases de la metodología y actividades.

4.1. Descripción y depuración de la base de datos

Se cuenta con una base de datos compuesta por 64554 secuencias genéticas provenientes del veneno de escorpión, proporcionadas por el grupo de investigación de farmacología de la Universidad del Valle. El archivo fuente se encuentra en formato de texto llamado *FASTA* (Figura **4-2**). Esta primera base de datos contiene todas las secuencias del transcriptoma (conocidas y desconocidas). Para filtrar las secuencias conocidas, el archivo fuente se comparó frente a la secuencias genéticas conocidas, disponibles en el repositorio GenBank, utilizando para el ello la herramienta de software duk.

```
>TRINITY_DN36600_c0_g1_i1 len=226 path=[0:0-225]
CCTTTGGCAGTTATTATTATTACAGCAGAAATCTGTGTGGGCATTATCTTGATTAACTTCTGTAACATTGCCACTTTTATATTGTACATT
TCTAATTCAGAACCCTGGCTAGCTTCTGGGGGCTGTTGGGATGAGTTGTTCTACTTCTGAGTTGGAAATTCAATTCATCCCAGAGGTGTT
CAATGGGATTAAGGTCTGGACTCTGGGCAGACTAGTTGAATCTTGC
>TRINITY_DN36572_c0_g1_i1 len=181 path=[0:0-180]
CTTGCATTACTTTAATACCTTGTAGAGCTTCATTTGTTCATAGCCTTTTTGCTCATTAATTCAATGACAGACAACCCTTGTCTATCAGTG
AATTAACTTAAAATTACATAATCATTAATATAATTTCACTATAGGTAAGATTTTATACACAAAATTTATGCATAGACTGGTTTTGTTTCT
T
>TRINITY_DN36553_c0_g1_i1 len=189 path=[0:0-188]
GCCCTCTTTAATAAATCTGGGCTCAGCTATTAGAACCCAGCAGAATGGTCCCTGCCTGGGGTTACAGGTGAATGTGAACTCATGTCAGTA
CTGGCTACTCTGGGGGACGGAAACAGTAACCCTCAGTACCATCTGAAGGCAGATGATTGTTGAAATTGGCATGGCTTCAGTATAGACTGC
TAAAAGCAA
>TRINITY_DN36508_c0_g1_i1 len=241 path=[0:0-240]
CTGTAGCAAAGTGAGGTTTCCGATCTTCCATTAATGTTCCTTCTTTTGAGCTGGTTTCTGTCAGTTCTACAATGATTGCTTCTTCTGGAA
TAAGCTGGGATTTAGCAGTTGACAAATTAATCTTTGCAATTTCTAAGTCATTTTCCTTGGTGCCTGTTTCAATTTCACAAATTGTAACAG
CTATGTTGTGAGGATGAACAGTTGAAGCTTTGAAACCTTTTGGCATATCATGAACTTGTAA
>TRINITY_DN36595_c0_g1_i1 len=287 path=[0:0-286]
TAGACAATGTTATCTGTGTTTCATTATCATTGGAAACATTCTGGACCTGTTTCTTCTTTGTAGGCCTTTTCCTTGCCTACTTTAGCATTC
AACTTATCAAGCAGGATTAGTATATCACTATGAGGTGTGGCAACAATGGTATTTTCAAGGCGGGTATAAAAGTCATCTTTCATTTCAGAC
TGACTGTCTTCTGTAGGGACATGAGCACTAAATGTTATGTTTTTTAATGCTGCCCATGATGTGAAGTCTGCAGATTCTGTTTTAATGGTT
TCAGAACTTATAAGGGC
```

Figura 4-2: Representación de las 5 secuencias genéticas en formato FASTA.

Luego de identificar las secuencias conocidas de la base de datos, estas se separaron de la base de datos obteniendo un total de 61951 secuencias desconocidas. En la Figura **4-3** se presenta de manera gráfica un ejemplo de las bases nitrogenadas, donde el color verde representa a la citosina (C), el color azul la timina (T), el color amarillo guanina (G) y el color rojo la adenina (A).

```
DNA vector of 61951 sequences
TRINITY_DN36600_c...  CCTTTGGCAGTTATTATTATTACAGCAGAAATCTGTGTGGGCATTATCTTGATTAACTTC... + 166 bases
TRINITY_DN36572_c...  CTTGCATTACTTTAATACCTTGTAGAGCTTCATTTGTTCATAGCCTTTTTGCTCATTAAT... + 121 bases
TRINITY_DN36553_c...  GCCCTCTTTAATAAATCTGGGCTCAGCTATTAGAACCCAGCAGAATGGTCCCTGCCTGGG... + 129 bases
TRINITY_DN36508_c...  CTGTAGCAAAGTGAGGTTTCCGATCTTCCATTAATGTTCCTTCTTTTGAGCTGGTTTCTG... + 181 bases
TRINITY_DN36595_c...  TAGACAATGTTATCTGTGTTTCATTATCATTGGAAACATTCTGGACCTGTTTCTTCTTTG... + 227 bases
TRINITY_DN36538_c...  TGGTTGAGTAACAGGCTATCTGCTGTTTTTTTCTGCTTACAATGGCTTTCAGAAAAAATT... + 333 bases
TRINITY_DN36527_c...  CGCAGAAACAGAAAAAAGGACAAAACTCCTGTGCCATAATGCCCTTCAATTCTTTTCAGA... + 398 bases
TRINITY_DN36534_c...  TCTTGAAATCGCAGTCGACTATCTGCACAAATTGCAAGAAAGTGCTCCGTATTCACCAGC... + 96 bases
TRINITY_DN36555_c...  CCTGGTAAGAATCTTTTGATATATCAAGAGCAAGTTCTAGTAATTTTCCAAATTTTTTAA... + 172 bases
TRINITY_DN36539_c...  CGAGTAGTGGCATCGGTCCGTCCCTTAAGCTCTTGCAGTTGAATGCACAGCGTTCTCGTT... + 114 bases
TRINITY_DN36568_c...  CAAAAATATTACCTGATAGCAACATTTGCACCAGTCACCTATATATGAAGATTGTAAATG... + 105 bases
TRINITY_DN36545_c...  AACGGAATTAGTTTTAGAAATGGACATCTGACGTGTATAGAATCCGAGGATGAGACGAAA... + 138 bases
... with 61939 more sequences.
```

Figura 4-3: Ejemplo de 12 secuencias genéticas de la base de datos.

4.2. Pre-procesamiento de datos

Se construyó una base de datos con 61951 registros y 21 variables: la primera variable es la longitud de cada secuencia y las 20 variables siguientes representan las distribuciones de frecuencia relativa de las k-tuplas para k= 1,2, donde la tabla K_1 hace corresponde a las 1-tupla (A, C, T, G), la tabla K_2 son las combinaciones de 2-tuplas (AA, AC, AT, AG,..., GG) (Figura **4-4**).

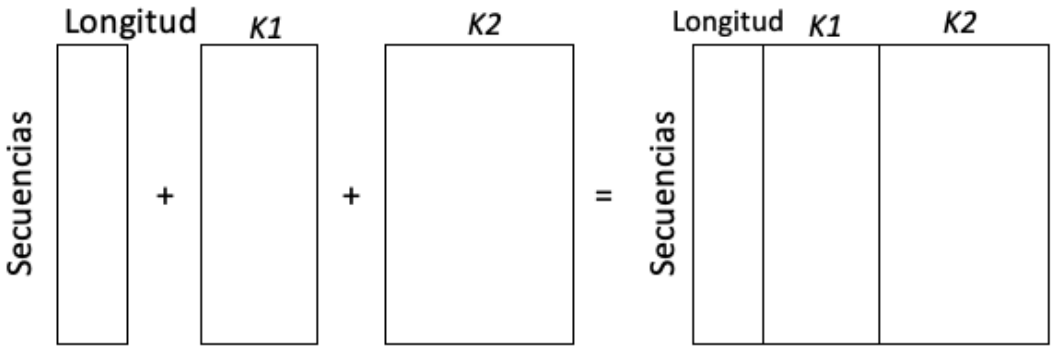

Figura 4-4: Estructura de la base de datos construida a partir de las subcadenas de las secuencias genéticas.

4.3. Análisis estadístico de las secuencias genéticas

En esta sección se presentan los métodos estadísticos utilizados para estudiar la base de datos construida, empezando con un análisis exploratorio, donde se muestra la distribuciones de frecuencia de las k-tuplas. Posteriormente, se utilizó el análisis multivariado para observar la relación entre las variables de las k-tuplas.

4.3.1. Análisis exploratorio de la base de datos

Para tener un conocimiento más amplio de las distribuciones de frecuencias de las k-tuplas de secuencias genéticas, se realizó un ejemplo seleccionando aleatoriamente dos secuencias genéticas de la base de datos para observar la distribución de frecuencia de las bases nitrogenadas, la distribución de frecuencia de las posiciones en las que se repite cada base nitrogenada en la secuencia genética, y mostrar la representación gráfica de las k-tuplas de las secuencias junto con la tabla de los aminoácidos. Con la base de datos construida con las k-tuplas de las secuencias genéticas, se observó la distribución de la longitud de todas las secuencias genéticas, la distribución relativa de la frecuencia en que se repite cada letra en la secuencia, la correlación entre 1-tupla, 2-tuplas y 3-tuplas.

4.3.2. Ejemplo de análisis de secuencias genéticas

Se tomó aleatoriamente 2 secuencias desconocidas de la base de datos (Figura **4-5**), con las que se construyó una base de datos con 86 variables, las cuales son las distribuciones de frecuencia relativa de las combinaciones hasta las 3-tuplas y la longitud de cada una de las secuencias.

```
[[1]]
[1] "GGTTACTGAGATGGACATAGTGTGCAAAGATCGTTGGATGTCATCATTCTCTAGTTCAATGTTTTATGTTGGAAGTCTTGGAGGAAGTTTATTTT
TTGGATATATGGGTGATAAGTATGGAAGACGTCCTGCATTTTTTGGCATGTTAATTGTCAACTTGGTCTTTGGCATTATGTCATGTTTTCCACCGAATTA
TATATGTTATATTATCTGCCGCATCATCTTAGGAACGTCGTATCCCTCTCTATTCCAGCTCCC"

[[2]]
[1] "TGAAAAAATGTTGATTATTGAGCAAAGGGCAAATATCAAATTTTGCCAAGGGTTAGGAAAGTCTACAACAGAAACTGCAGTTGCTGTACCAAGTT
TCCAAGTTTATGGTGATCAAGCCCTGAACCGAATCTCTGTATTTGAGTGGCACCGTCGTTTCAAGGAAGGCAGAGAGAAAACTCAAGATGGCGAACATAC
CAGACGGCCAGTTACCGTTCGAACTGAAGCCACTACATTCTGACCGTGCAAAAGCTAGTA"
```

Figura 4-5: Dos secuencias aleatorias de la base de datos.

4.3.3. Distribución de frecuencia de las bases nitrogenadas según su aparición.

Para encontrar un patrón en la repetición de cada base en las dos secuencias anteriormente tomadas de forma aleatoria, se calculó la distribución de frecuencia, donde se tuvo en cuenta

la posición en la que aparecía la base por primera vez y la posición en la que volvía aparecer. En el caso de la secuencia 1 (Figura **4-6**), se observa las diferencias en las distribuciones relativas de cada base, debido a que la cantidad de repeticiones o apariciones en la secuencia son distinta para cada base, como la base T que al inicio de la secuencia es la base que más demora en volver se repetir pero a partir de la segunda aparición se repite más frecuente por lo que el valor de la frecuencia es bajo, al igual que con la base C. Mientras, en la base A, su frecuencia de reincidencia en la secuencia no muy seguido al comienzo de la secuencia, aunque fue aumentando su repetición lo mismo que con la base G.

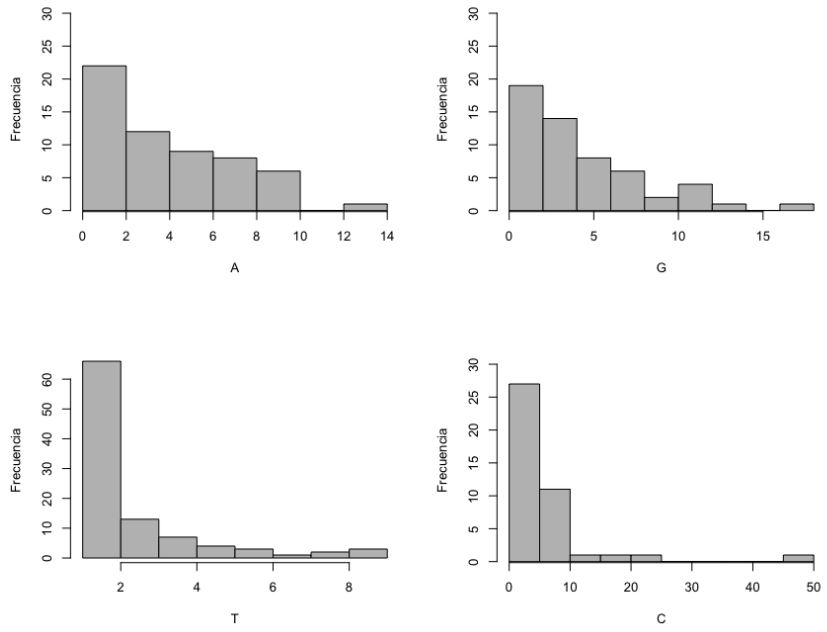

Figura 4-6: Distribución de la frecuencia relativa de 1-tupla de la secuencia 1.

En la secuencia 2 (Figura **4-7**), se muestra las distribuciones relativas de cada base nitrogenada (A, T, C, G), en cual se observa que las distribuciones de las bases G y C hay una alta frecuencia de apariciones seguidas a lo largo de la secuencia genética, caso contrario son las bases A y T, ya que ellas presentan baja frecuencia cuando aumenta la repetición de la bases nitrogenadas en la secuencia genética.

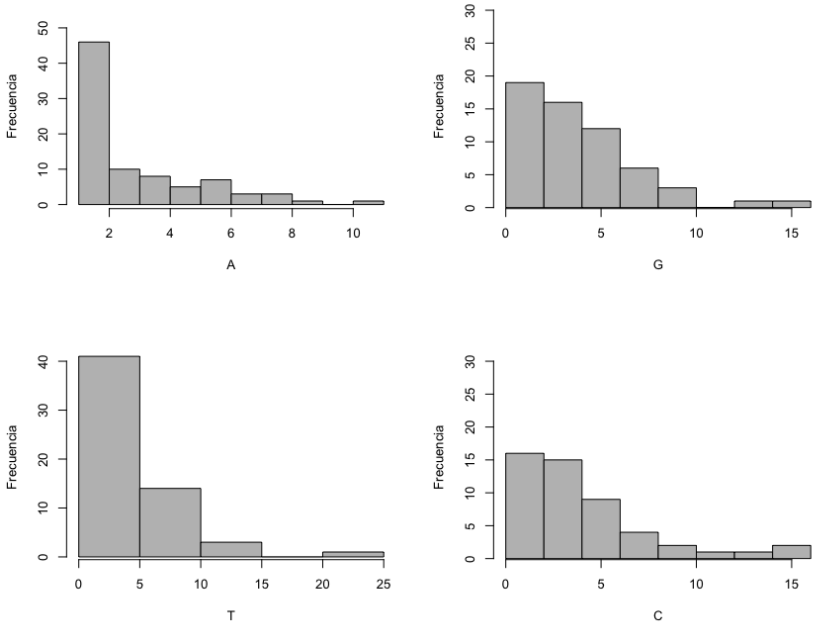

Figura 4-7: Distribución de la frecuencia relativa de 1-tupla de la secuencia 2.

4.3.4. Representación gráfica de las secuencias de acuerdo con la tabla de aminoácidos

En las secuencias genéticas, las combinaciones de 3-tuplas se traducen en el lenguaje de las proteínas, identificando los aminoácido que conforman las proteínas, según la combinación de bases nitrogenadas que se tiene en cada secuencia. Para ambas secuencias se determinó en porcentaje la cantidad de 3-tuplas. En la secuencia 1 (Figura **4-8**) se encontró que los aminoácidos más frecuentes en esta secuencia son Fenilalanina (Phenylalanine en inglés) con un 7.03 %, Leucina (Leucine en inglés) con un 6.64 % , y aunque la frecuencia de los demás animoácidos es baja, el aminoácido Alanina (Alanine en inglés) es el menos frecuente. En cambio, en la secuencia 2 (Figura **4-9**) el aminoácido con más apariciones es Lisina (Lysine en inglés) con 9.49 %, sin embargo el animoácido con menor frecuencia en la secuencia es Triptófano (Tryptophan en inglés).

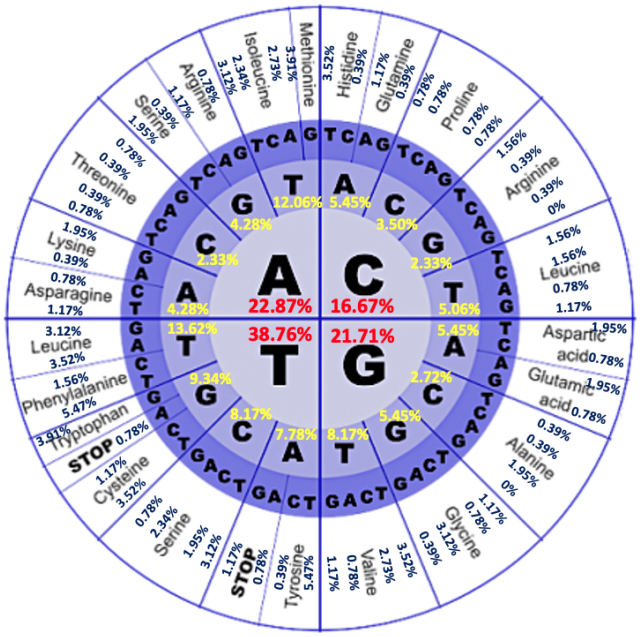

Figura 4-8: Diagrama circular de los aminoácidos según la distribución de frecuencia relativa de las k-tuplas de la secuencia 1.

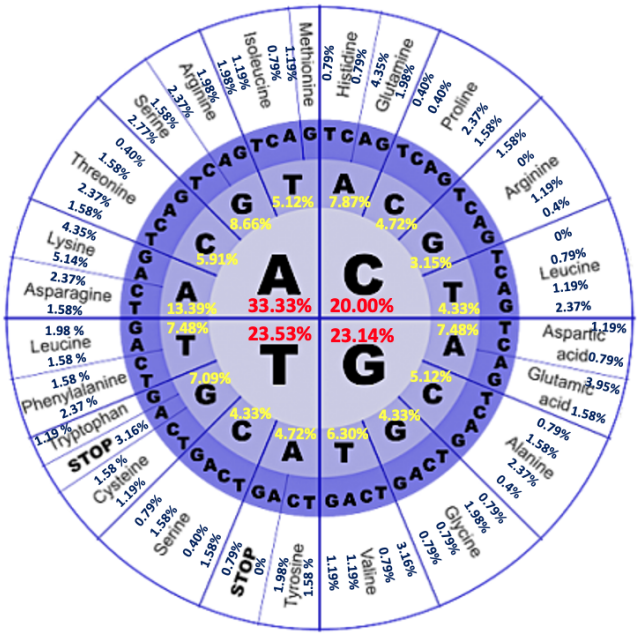

Figura 4-9: Diagrama circular de los aminoácidos según la distribución de frecuencia relativa de las k-tuplas de la secuencia 2.

4.4. Construcción de la base de datos

Para el análisis estadístico de todas las secuencias desconocidas se construyó la base de datos compuesta de la variable longitud de las secuencias, las distribuciones de frecuencia relativa para 1-tupla y para 2-tuplas como se muestra en la Figura **4-10**, obteniendo un total de 21 variables. Esta base de datos es la que se utilizó para el análisis multivariado como el ACP y el análisis de conglomerados.

	Longitud	A	C	G	T	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	226	0.2301	0.1770	0.2301	0.3628	0.0533	0.0400	0.0533	0.0844	0.0578	0.0267	0.0000	0.0889	0.0533	0.0400	0.0800	0.0578	0.0667	0.0667	0.0978	0.1333
2	181	0.3149	0.1657	0.1105	0.4088	0.0889	0.0611	0.0444	0.1222	0.0778	0.0222	0.0000	0.0667	0.0333	0.0278	0.0111	0.0389	0.1167	0.0500	0.0556	0.1833
3	189	0.2804	0.2169	0.2540	0.2487	0.0904	0.0532	0.0691	0.0638	0.0691	0.0532	0.0053	0.0904	0.0585	0.0585	0.0851	0.0532	0.0638	0.0532	0.0904	0.0426
4	241	0.2490	0.1701	0.1950	0.3859	0.0875	0.0292	0.0542	0.0750	0.0625	0.0250	0.0042	0.0792	0.0542	0.0417	0.0333	0.0667	0.0458	0.0708	0.1042	0.1667
5	287	0.2509	0.1707	0.2021	0.3763	0.0629	0.0420	0.0559	0.0909	0.0699	0.0210	0.0035	0.0734	0.0455	0.0385	0.0524	0.0664	0.0734	0.0699	0.0909	0.1434
6	393	0.2977	0.2366	0.1705	0.2952	0.1276	0.0714	0.0408	0.0587	0.0969	0.0510	0.0255	0.0638	0.0306	0.0383	0.0536	0.0459	0.0434	0.0765	0.0510	0.1250
7	458	0.2904	0.1900	0.1616	0.3581	0.0985	0.0460	0.0460	0.1007	0.0744	0.0328	0.0131	0.0700	0.0503	0.0328	0.0328	0.0460	0.0678	0.0766	0.0700	0.1422
8	156	0.2949	0.2115	0.2372	0.2564	0.0903	0.0710	0.0581	0.0774	0.0903	0.0129	0.0516	0.0581	0.0645	0.0645	0.0452	0.0581	0.0516	0.0645	0.0839	0.0581
9	232	0.3190	0.1336	0.2069	0.3405	0.0909	0.0260	0.0866	0.1169	0.0433	0.0390	0.0087	0.0390	0.0996	0.0216	0.0346	0.0519	0.0866	0.0433	0.0779	0.1342
10	174	0.2184	0.2299	0.2701	0.2816	0.0520	0.0289	0.0867	0.0520	0.0578	0.0405	0.0578	0.0751	0.0462	0.0694	0.0578	0.0925	0.0636	0.0867	0.0694	0.0636
11	165	0.3939	0.1636	0.1697	0.2727	0.1402	0.0549	0.0793	0.1220	0.0854	0.0244	0.0061	0.0427	0.0976	0.0427	0.0122	0.0183	0.0732	0.0366	0.0732	0.0915

Showing 1 to 11 of 61,951 entries, 21 total columns

Figura 4-10: Ejemplo de 10 secuencias de la base de datos con las variables construidas.

4.5. Análisis multivariado de las secuencias genéticas

Se evaluó las secuencias genéticas sobre la distribución relativa de las letras de 1-tupla y de 2-tuplas, a través del ACP con posibles componentes que permitan encontrar similitudes entre las secuencias. Este método que reduce el número de variables, obteniendo las nuevas componentes principales que son las combinaciones lineales de las variables de la base de datos, utilizando los paquetes FactoMineR y FactoClass del software R (Lê et al. 2008) y (Pardo & Del Campo 2007). Además, se le agregó al ACP la variable de longitud como variable suplementaria, para mostrar la relación con las variables de las k-tuplas.

4.5.1. Agrupación de las secuencias genéticas

Se realizó un análisis de conglomerado con los resultados del ACP para agrupar las secuencias, teniendo en cuenta la distancia entre las k-tuplas que es la distancia euclídea, ya que de acuerdo con los antecedentes es la distancia más empleada en estos análisis. Además, se estimó el número óptimo de cluster con el método de codo, en el cual se determinó que 4 grupos es una buena opción este estudio. Es importante tener en cuenta, que dado el volumen de datos que se maneja en este trabajo, el algoritmo de k-means tiene muchas iteraciones, lo

cual hace que se pierda la dependencia al valor semilla inicial. Además, al realizar una sola agrupación de este número alto de secuencias genéticas se tiene un alto costo computacional.

4.5.2. Caracterización de los grupos de las secuencias genéticas

Al agrupar las secuencias genéticas, se estudia las características distintivas en los grupos, por medio de las k-tuplas para identificar patrones o comportamientos de las distribuciones de la frecuencia de estas variables. Luego, se asocia las distribuciones de frecuencia de las k-tuplas al diagrama de aminoácidos para conocer los aminoácidos más y menos representativos en cada grupo.

5 Resultados

En este capítulo se presentan los resultados obtenidos en este trabajo, estudiando las variables construidas a partir de las secuencias genéticas desconocidas de la glándula de veneno de escorpión. En el cual, se presenta un análisis exploratorio de todas las secuencias genéticas desconocidas y un análisis multivariado, en el cual se conozca la relación entre las variables y se identifique similitudes entre las secuencias.

5.1. Análisis exploratorio de las secuencias genéticas

En esta sección se presenta la distribución de las variables construidas, es decir tener una vista general del comportamiento de las variables: 1-tupla, 2-tuplas, 3-tuplas y la longitud de las secuencias, además se estudió la correlación entre las variables.

La distribución de la variable longitud de las secuencias genéticas (Figura **5-1**), presenta una asimetría positiva, dado que la mayoría de las secuencias tienen una longitud entre 100 a 300, mientras muy pocas secuencias tienen una longitud muy largas.

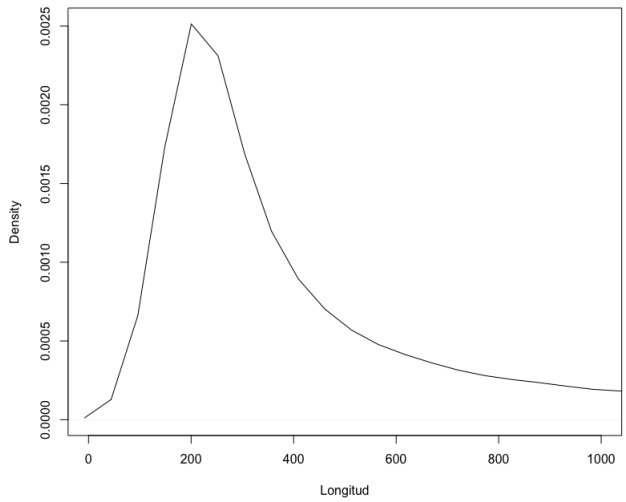

Figura 5-1: Distribución de la longitud de las secuencias genéticas de la base de datos.

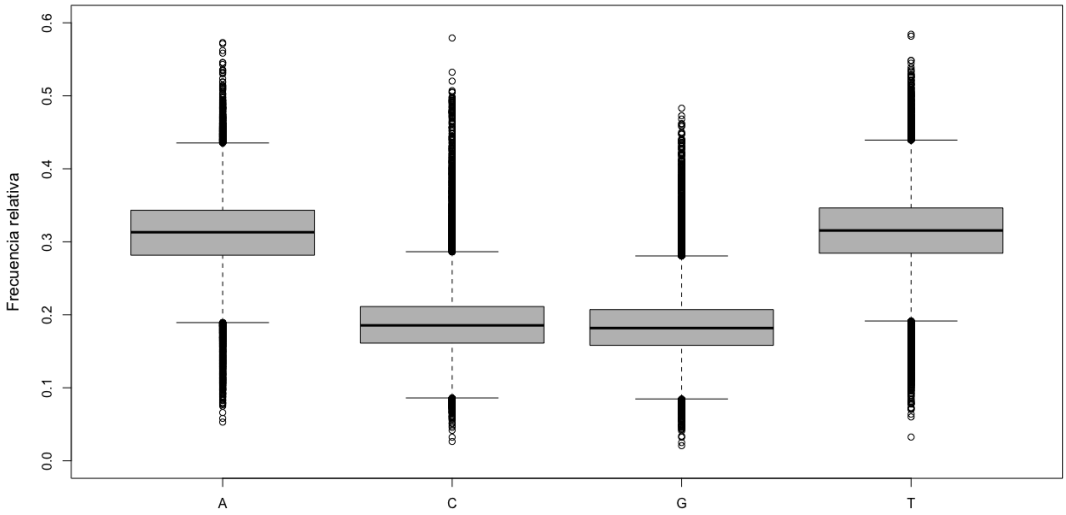

Figura 5-2: Distribución de la frecuencia relativa de 1-tupla de la base de datos.

En la Figura **5-2**, se muestra la distribución de 1-tupla de las variables, en la cual se aprecia que las bases A y T presentan similar frecuencia siendo las que más aparecen, mientras las bases G y C son las que con menos frecuencia se encuentran en las secuencias y al igual que con el caso anterior estas presentan valores muy similares, esto se puede deber a que en las secuencias genéticas de ADN cuando se tiene la doble hélice, es decir la unión de secuencias genéticas que se complementan, la base A se conecta con la base T y la base G con C.

Para el estudio del comportamiento de las variables con 2-tuplas (Figura **5-3**), se observa que las combinaciones que más se encontraron en las secuencias son las bases AA y TT, aunque la combinación de CC es la variable que presenta mayor cantidad de puntos atípicos, que se puede deber porque en algunas secuencias hay presencia de aminoácidos que requieran de esta combinación. En cambio, la variable con la combinación CG es la que presenta menor frecuencia en las secuencias, seguido de las variables GG y CC.

En el caso de las variables con 3-tuplas (Figura **5-4**), se observa que las combinaciones menos frecuentes en las secuencias genéticas que contienen, sea al inicio o al final, la combinación CG como lo es en el caso ACG, CCG, CGC, CGA, CGG y CGT, seguida de la combinación CCC que es la variable más valores atípicos. Mientras, la variable con mayor frecuencia en las secuencias genéticas es la combinación AAA, y en general las combinaciones que contengan AA tienden tener una mayor frecuencia en apariciones en las secuencias.

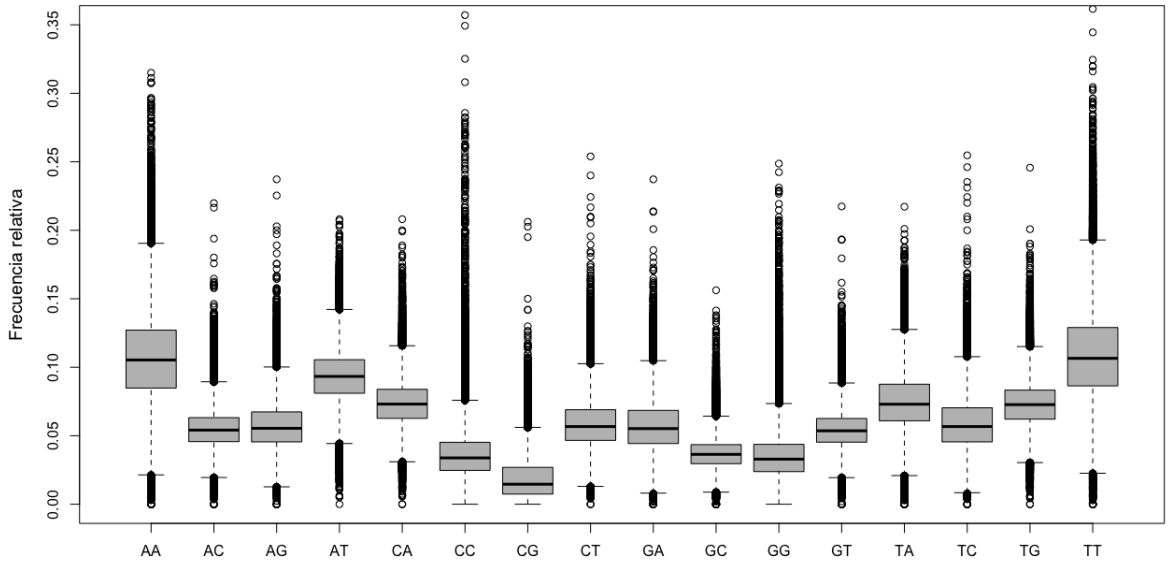

Figura 5-3: Distribución de la frecuencia relativa de 2-tuplas de la base de datos.

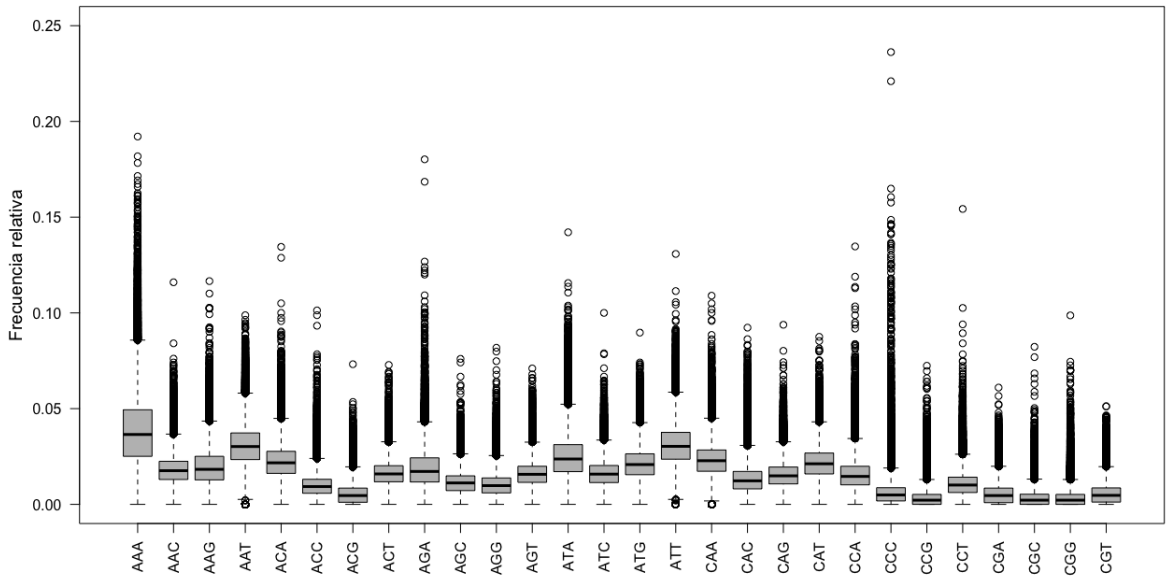

Figura 5-4: Distribución de la frecuencia relativa de 3-tuplas de la base de datos.

Después de evaluar el comportamiento de cada una de las variables de la base de datos, es importante conocer la relación entre la distribución de frecuencias de las k-tuplas.

En la Figura **5-5**, se presenta la correlación de las variables con 1-tupla, en la cual permite observar que las bases A y C tienen una correlación negativa, así como las bases G y T. Mientras que, las bases que han obtenido una correlación muy cercana al 0 ha sido C y G.

En el caso de las variables de 2-tuplas (Figura **5-6**), se observa una correlación positiva fuerte entre las variables que contienen una combinación similar como: AC con CA, AG con GA, AT con TA, CT con TC y GT con TG, en cambio algunas variables presentan una correlación negativa fuerte, como lo son: AA con CT, AA con TC, AG con TT, AT con CC, AT con CG, AT con GC, GA con TT y GC con TA, debido a la correlación de 1-tupla que algunas letras presentan correlación negativa. Sin embargo, al estudiar la correlación en variables de 3-tuplas (Figura **5-7**), se muestra que no son muy explicativas, ya que la relación entre ellas esta condicionada por la correlación que existan en las 2-tuplas, por lo que para el análisis multivariado se construyó la base de datos hasta con 2-tuplas.

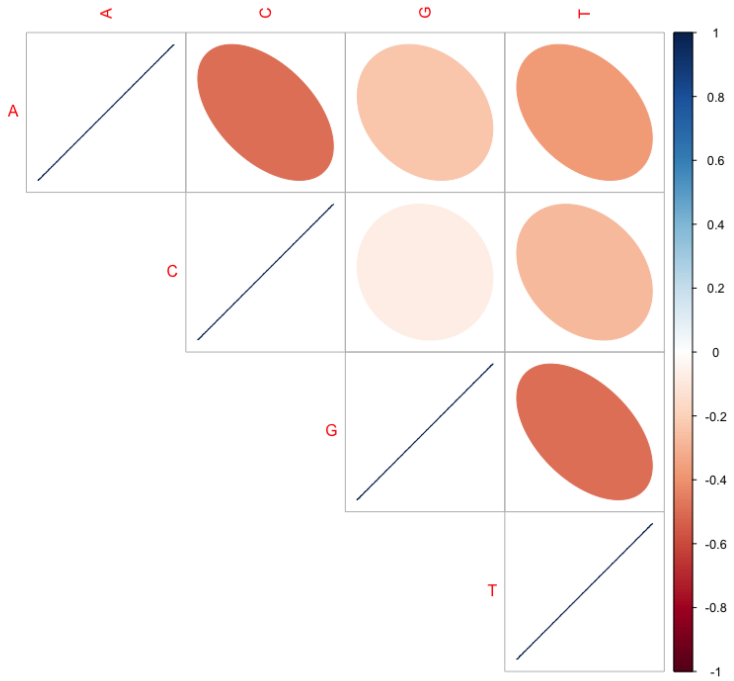

Figura 5-5: Correlación entre las variables de 1-tupla de la base de datos.

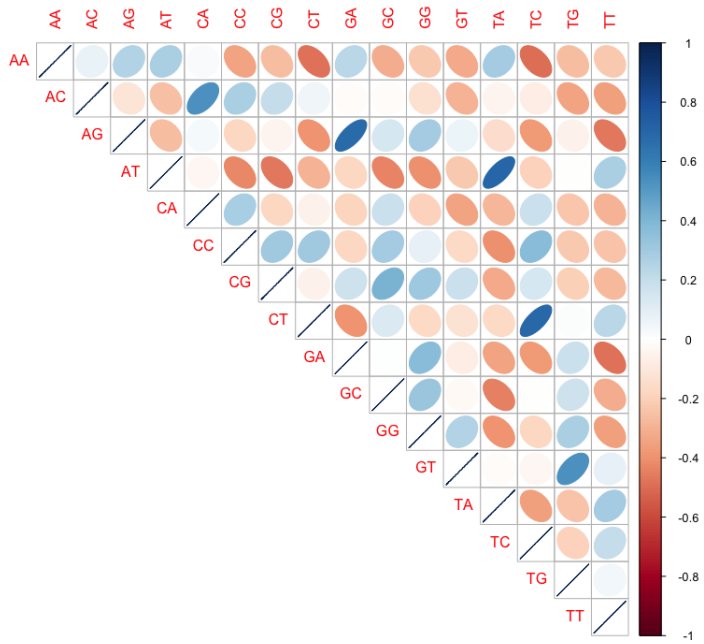

Figura 5-6: Correlación entre las variables de 2-tuplas de la base de datos.

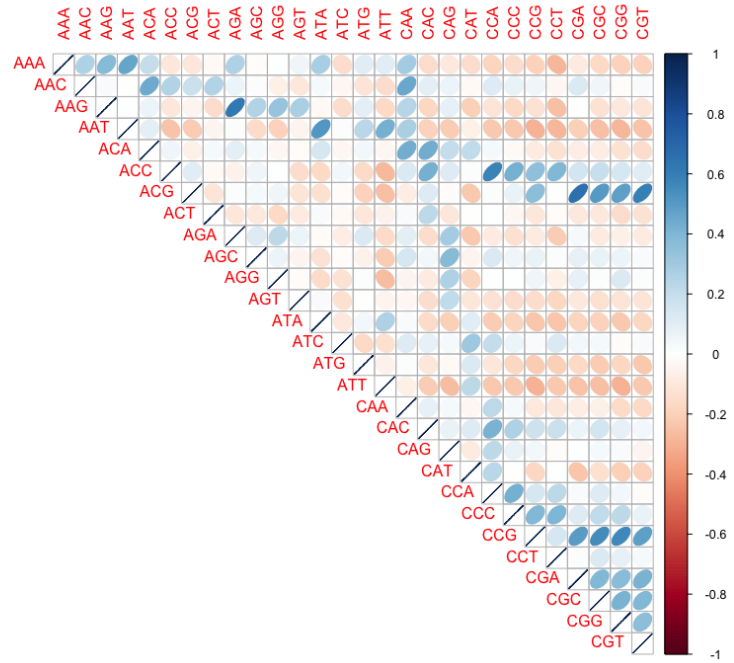

Figura 5-7: Correlación entre las variables de 3-tuplas de la base de datos.

5.2. Análisis multivariado de las secuencias genéticas

5.2.1. Análisis de Componentes Principales (ACP)

Los valores propios (λ_q) proporcionan directamente las inercias proyectadas sobre cada una de las direcciones que se busca. El comportamiento de estos presentan un decrecimiento (Tabla **5-1**), pues se ordenan según la variabilidad explicada, por está razón, se observa un valor sumamente alto para la primera componente, ya que será quien explica el mayor porcentaje de variabilidad presente en los datos.

Se calculó la inercia de los componentes (Tabla **5-1**), esto tomando cada valor propio y dividiéndolo entre el número de variables estudiadas, continuamente se multiplica por 100, de esta forma obtener el porcentaje de variabilidad explicado por cada eje, es posible evidenciar que la primera componente explica cerca del 25.8296 % de la inercia, seguidamente la componente 2 obtiene un 23.5807 %, esto puede revelar aspectos sistemáticos de los datos, lo que implica que el primer plano factorial conserva el 49.410 % de la inercia total (Figura **5-8**).

Si bien la componente 3 tiene un valor significativo de porcentaje de inercia (Tabla **5-1**), se decidió por medio del criterio de consistencia de los resultados que lo más conveniente era tomar las 2 primeras componentes, ya que al momento de realizar la agrupación de las secuencias con la tercera componente presentaban zonas de solapamiento en la representación gráfica, es decir zonas en las que las secuencias estaban muy unidas o encima de otras dificultando la calidad de los resultados. Mientras con 2 componentes, los grupos fueron más separados permitiendo tener una agrupación más consistente como se observa en la siguiente subsección 5.2.2.

Componente	**Valor Propio**	**Inercia (%)**	**Inercia Acumulada (%)**
1	5.1659	25.8296	25.8296
2	4.7161	23.5807	49.4103
3	3.4920	17.4602	66.8705
4	1.2122	6.0611	72.9316
5	1.0717	5.3583	78.2899
6	0.8499	4.2493	82.5392
7	0.7024	3.5120	86.0512
8	0.6553	3.2764	89.3276
9	0.5990	2.9948	92.3225
10	0.5473	2.7365	95.0589

Tabla 5-1: Valor propios y porcentaje de varianza explicada.

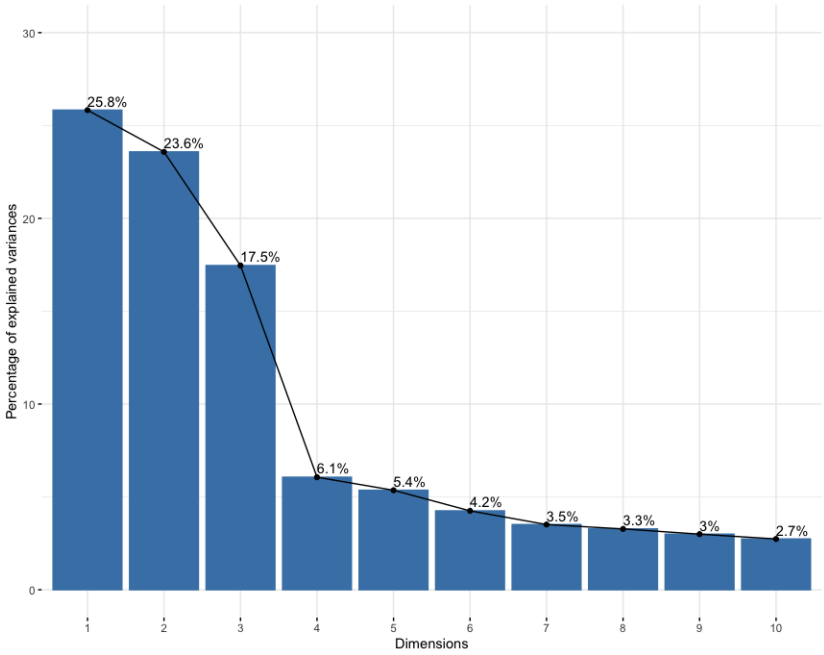

Figura 5-8: Distribución de la varianza explicada del ACP de la base de datos.

Por medio del gráfico de contribuciones de las variables en los ejes 1 y 2 (Figura **5-9**), se resalta que las variables que más contribuyen en la primera y segunda componente principal son las variables de 1-tupla (color rojo), además se observa que base A y la base C tienen una correlación muy baja como se había visto en el análisis exploratorio, al igual que la base G con la base T, por lo que las combinaciones de AC, CA, GT y TG tienen poca contribución en los ejes. Ahora bien, las 2-tuplas de AG, GA, AA, TT, GT, TC y CC son variables que más contribuyen en la formación de los ejes, ya están mas alejadas del punto de origen.

5.2.2. Agrupación de las secuencias genéticas

Antes de realizar la agrupación por el método de k-means, es necesario calcular el número óptimo de grupos o cluster que se van a formar cuando no se tiene información sobre cuántos grupos se requieren; este número se puede determina, utilizando el algoritmo de k-means en los datos y se analiza la suma total de cuadrados internos como se muestra en la Figura **5-10**, en el cual a partir de 4 grupos se observa que los valores tienden a ser constantes, lo que indica que agrupar las secuencias en 4 grupos es adecuado.

En la Figura **5-11**, se observa los 4 grupos en los cuales se agruparon las secuencias genéticas a partir del ACP realizado anteriormente, permitiendo conocer las variables que aportan variabilidad en las secuencias genéticas.

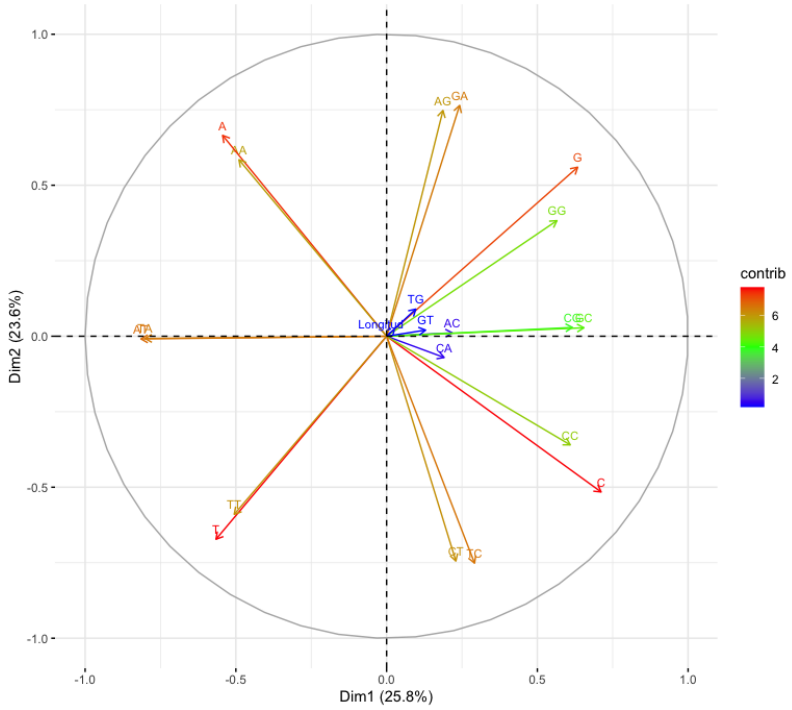

Figura 5-9: Circulo de Correlación y Contribuciones en ACP.

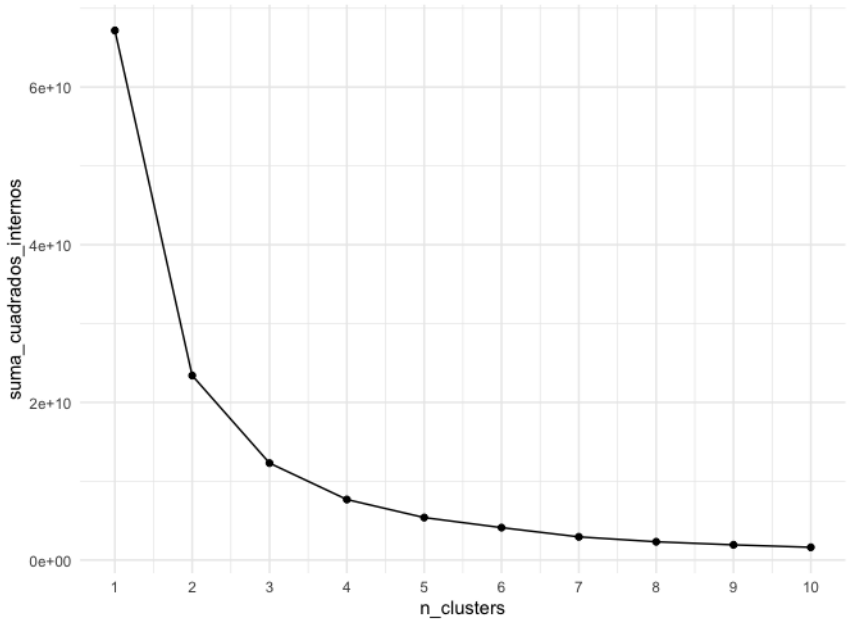

Figura 5-10: Número óptimo para agrupar las secuencias genéticas.

Figura 5-11: Agrupación de las secuencias genéticas por el algoritmo k-means.

En la relación entre las variables con los 4 grupos de las secuencias genéticas, los cuales se observan en la Figura **5-12**, se resalta que los grupos tienen mayor relación con una de las componentes, por ende contienen mas frecuencia de las k-tuplas asociadas a esa componente, esto es debido a que la composición de la secuencia genética, ya que como se ha visto en el análisis exploratorio la distribución de las k-tuplas puede variar.

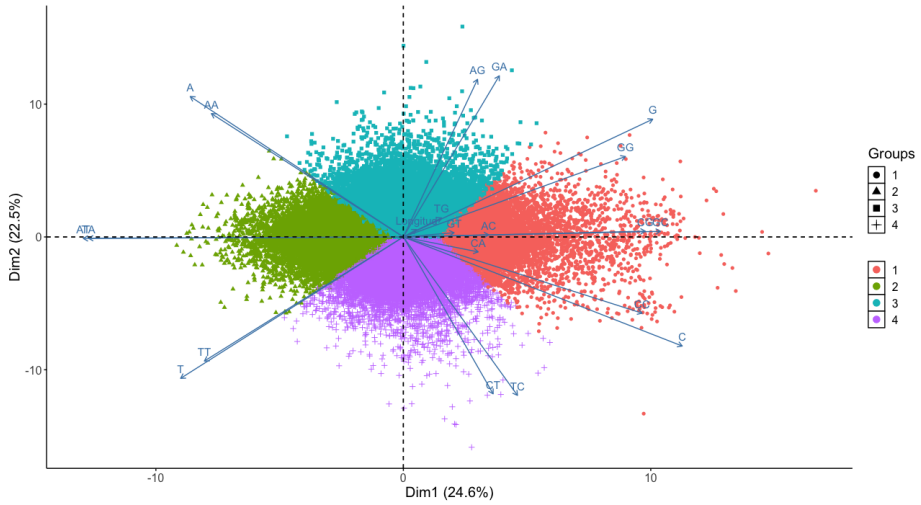

Figura 5-12: Representación simultánea de las secuencias genéticas y las variables.

5.2.3. Caracterización de los grupos de secuencias

Al identificar similitudes entre secuencias y clasificarlas en 4 grupos, se realizó un análisis descriptivo en los grupos encontrando características en las distribuciones de frecuencias de las k-tuplas. En la Figura **5-13**, se observa la distribución de las secuencias de la base de datos según los grupos, en el cual el Grupo 4 contiene el 32 % de las secuencias, seguido del Grupo 3 que tiene el 31 % de las secuencias, el Grupo 1 solo tiene el 20 % de las secuencias y el grupo con menos secuencias es el Grupo 2, en el cual solo hacen parte el 13 % de las secuencias.

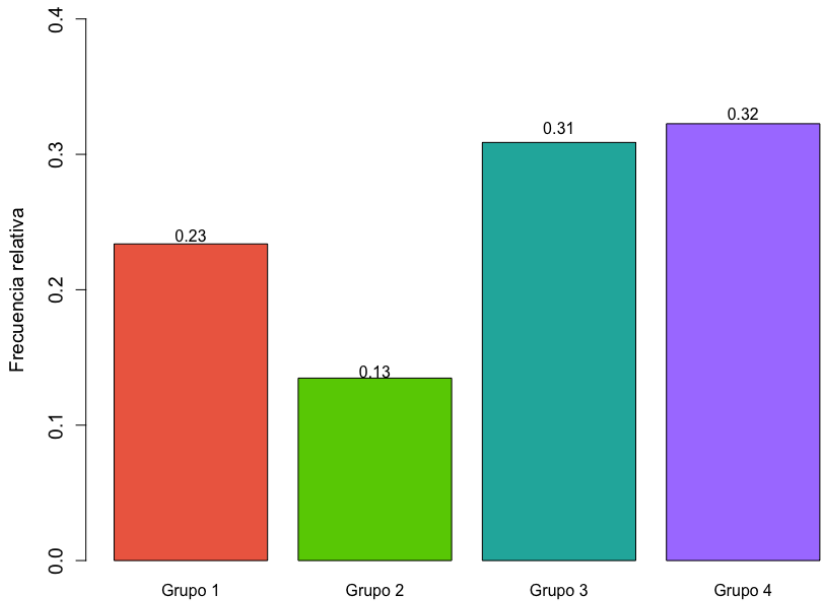

Figura 5-13: Distribución de frecuencia relativa de las secuencias genéticas en los 4 grupos.

Al analizar la distribución de 1-tupla en cada grupo (Figura **5-14**) se encontraron diferencias, ya que se observó que en el Grupo 1, hay más presencia de las bases nitrogenadas A y T que las bases C y G. En el Grupo 2, la distribución de las bases son más similares, aunque predominan las bases A y T. Los grupos 3 y 4 presentan una característica similar y es que una de sus bases es la es más frecuente en las secuencias de estos grupos; en el Grupo 3, la base T es la más frecuente y la base G es la menos frecuente, mientras en el Grupo 4, la bases que se presencia tiene en las secuencias es la A y la que menos aparece es la base C.

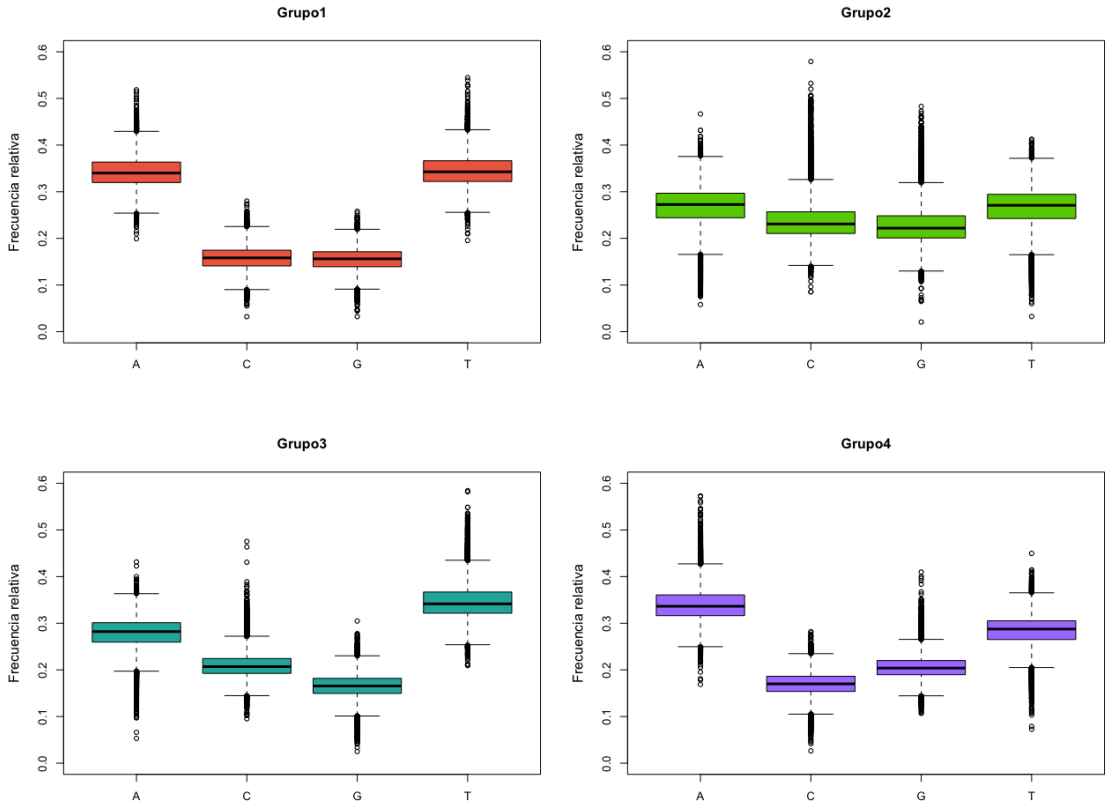

Figura 5-14: Distribución de la frecuencia relativa de 1-tupla para los 4 grupos.

En el caso de las distribuciones de 2-tuplas (Figura **5-15**), se encontró que en el Grupo 1, que las tuplas que tienen la base A y T son las que tienen mayor frecuencia y las pocas frecuentes con las tuplas es las cuales se tiene las combinaciones CG, CC, GC y GG. En el Grupo 2, se observó que las variables no presentan una variabilidad alta, aunque la variable con la combinación de CG es la que menos presencia tiene. El Grupo 3, contiene con más frecuencia las combinaciones en las cuales la base T se encuentre y con poca ocurrencia con la base G, como es el caso de la combinación CG que su presencia es muy baja. En cambio, en el Grupo 4 las variables que contiene la base A son las frecuentes y las variables con la base C son poco repetidas, al igual que con el Grupo 3, la variable CG es la común en este grupo.

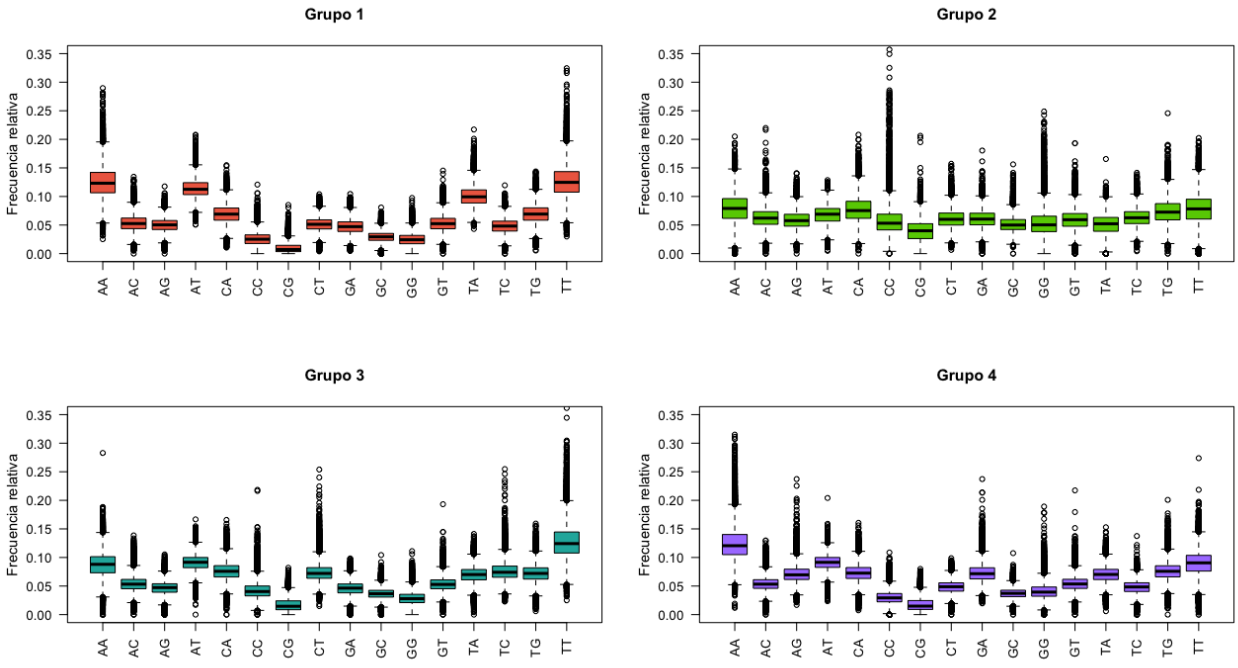

Figura 5-15: Distribución de la frecuencia relativa de 2-tupla para los 4 grupos.

En el comportamiento de las variables con 3-tuplas, se observó en la Figura **5-16** que la variabilidad entre las variables en los 4 grupos fueron disminuyendo, sin embargo en el Grupo 1 se encontró una variabilidad más alta en las variables que sus combinaciones predominaban las bases A y T como es el caso de las variables AAA y TTT. El Grupo 2 es el que menos variabilidad presenta en sus variables, ya que como se observó en la distribución de 1-tupla, la cantidad de bases en este grupo son muy cercanos. En los grupos 3 y 4, se presenta un comportamiento inverso entre ellos, ya que las combinaciones con la base T son las que presentan mayor variabilidad y en el grupo 4 la base A es la que influencia a una mayor variabilidad en las variables.

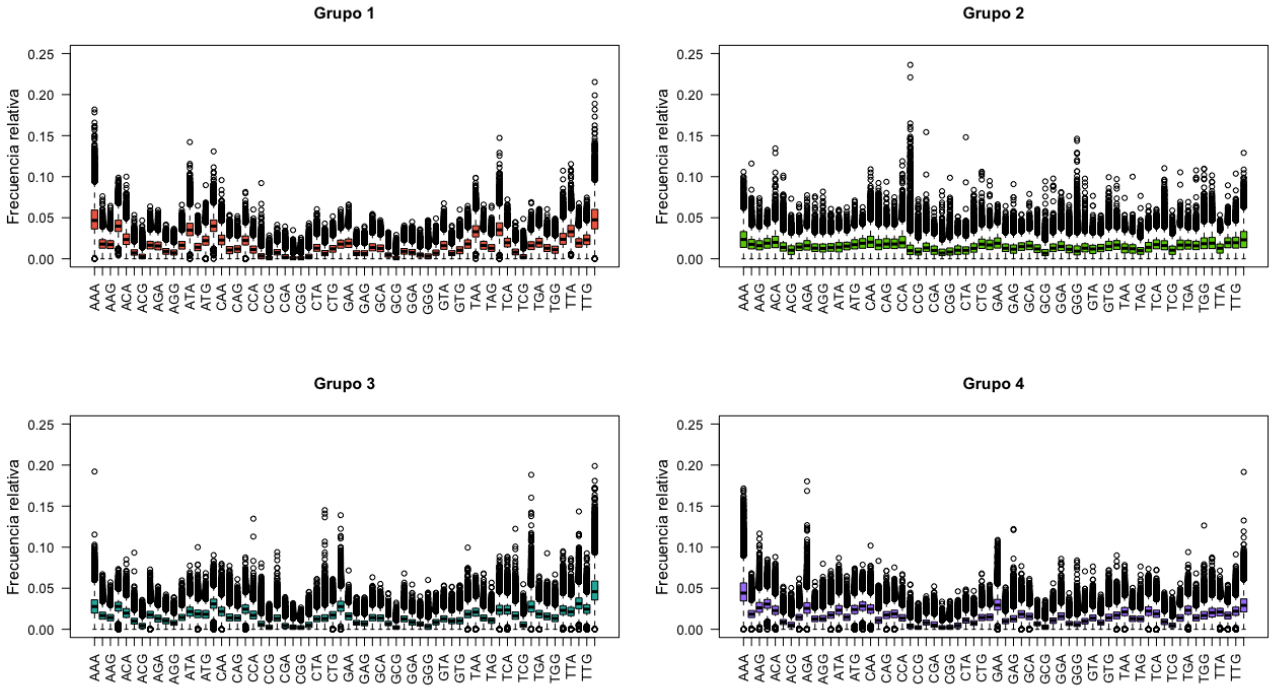

Figura 5-16: Distribución de la frecuencia relativa de 3-tupla para los 4 grupos.

Representación gráfica de los grupos de acuerdo con la tabla de aminoácidos

Al conocer la distribución de las combinaciones de 3-tuplas de las bases nitrogenadas, se identificó el porcentaje de presencia de cada aminoácidos para los grupo. En la Figura **5-17**, se observa de la distribución en términos porcentual de 1-tupla, 2-tuplas y 3-tuplas de las bases, además se puede conocer que en este grupo los aminoácidos con mayor presencia son *Fenilalanina*, *Lisina* y *Isoleucina*, mientras este grupo de secuencias presenta ausencia del aminoácido *Arginina*.

El Grupo 2 las secuencias se caracterizan por tener presencia de todos los aminoácidos de la tabla como se observa en la Figura **5-18**, sin embargo, los aminoácidos que más se destacan en este grupo son *Serina*, *Leucina*, *Treonina* y *Valina*. En cambio, en el Grupo 3 los aminoácidos más frecuentes son *Fenilalanina*, *Isoleucina*, *Serina* y *Asparagina*, y con una presencia muy baja de los aminoácidos *Arginina*, *Triptófano* y *Metionina*. Por último, en el Grupo 4 se caracteriza por tener una mayor cantidad de aminoácidos como *Lisina*, *Asparagina*, *Isoleucina* y *Fenilalanina*, por el contrario, los aminoácidos *Arginina*, *Prolina*, *Metionina* y *Triptófano* son muy escazos en este grupo de secuencias.

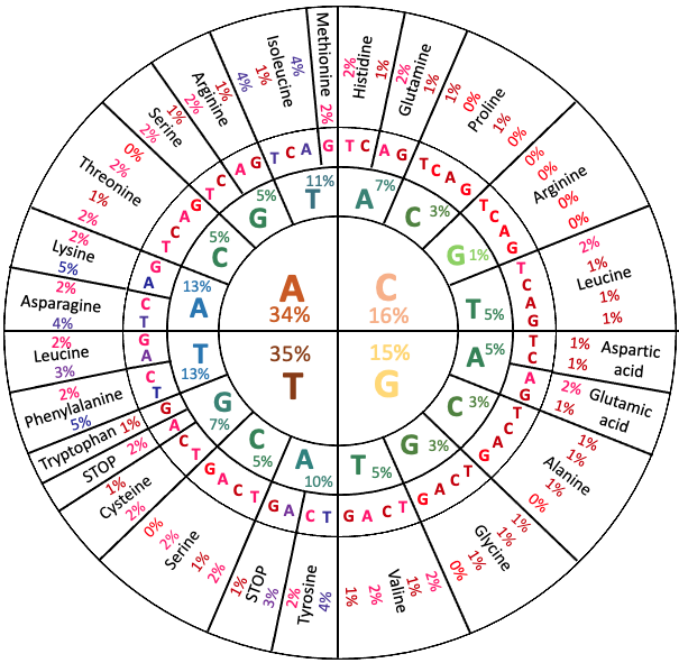

Figura 5-17: Diagrama circular de los aminoácidos según la distribución de frecuencia relativa de las k-tuplas del Grupo 1.

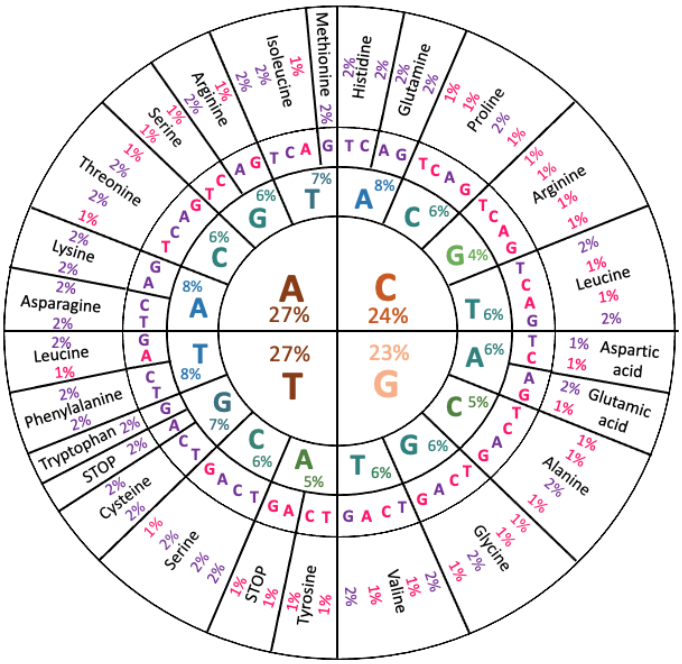

Figura 5-18: Diagrama circular de los aminoácidos según la distribución de frecuencia relativa de las k-tuplas del Grupo 2.

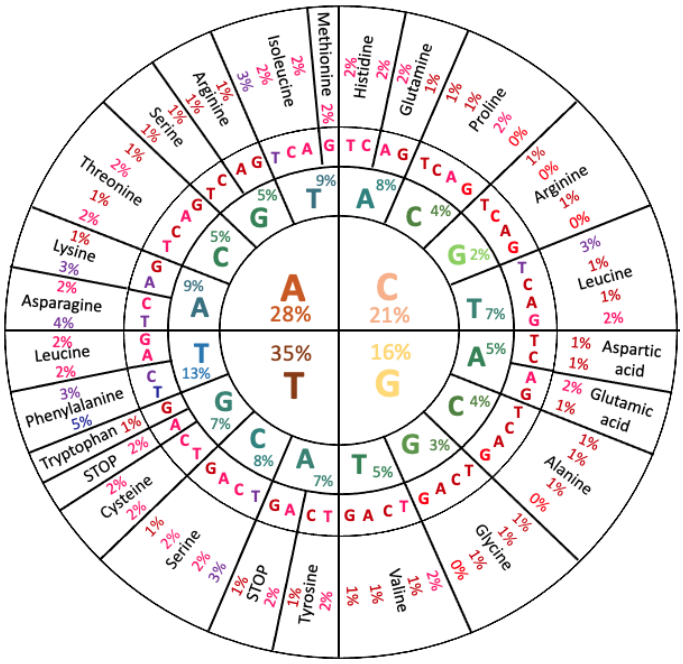

Figura 5-19: Diagrama circular de los aminoácidos según la distribución de frecuencia relativa de las k-tuplas del Grupo 3.

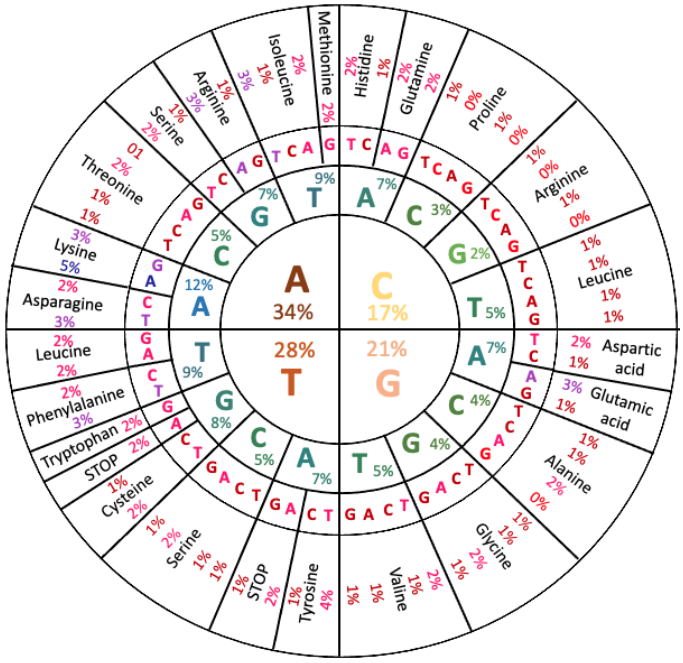

Figura 5-20: Diagrama circular de los aminoácidos según la distribución de frecuencia relativa de las k-tuplas del Grupo 4.

6 Conclusiones y recomendaciones

6.1. Conclusiones

En este trabajo se desarrolló el análisis de secuencias genéticas desconocidas provenientes de la glándula de veneno del escorpión, en las cuales no se tuvo en cuenta el marco abierto de lectura y péptido de señal con el fin de centrar el estudio en la búsqueda de aminoácidos y no en la formación de proteínas. El análisis de estas secuencias genéticas permitió conocer la relación que existe en la composición de las secuencias, por medio de las distribuciones de las k-tuplas con el diagrama de los aminoácidos.

En el análisis de componentes principales (ACP) se evidenció que las variables de 1-tupla, aparentemente las bases A y T disminuyen cuando las bases C y G aumentan, debido a que las bases C y G se encuentran altamente correlacionadas, al igual que las bases A y T. También, en las variables de 2-tuplas se encontró que las tuplas AT, TA, CG y GC determinan la primera componente, mientras la segunda componente la determina las tuplas AC, AG, CA, CT, TC, TG , GA y GT, debido que hay una relación que existen en las bases nitrogenadas cuando se conectan dos secuencias genéticas y forman la doble hélice, donde la base A se conecta con la base T y la base C se conecta con la base G.

En el análisis de conglomerados con el algoritmo de k-means se determinó que el número óptimo de grupos para estas secuencias genéticas desconocidas son 4. Estos grupos se conformaron de acuerdo a su similaridad en las distribuciones de k-tuplas y los resultados del ACP, permitiendo realizar una caracterización de los grupos e identificar los aminoácidos más influyentes en cada grupo, encontrando una característica distintiva entre ellos.

6.2. Recomendaciones

Un análisis interesante puede ser comparar los aminoácidos identificados en cada grupo con los aminoácidos que conforman las proteínas que se encuentran registradas en las bases de datos proteínas internacional como Protein Data Bank (PDB), con el fin de identificar si en estos grupos hay presencia de información para la producción de proteínas y si hay alguna similitud con las secuencias genéticas de otros venenos.

Para trabajos futuros podría ser interesante desarrollar el análisis de secuencias genéticas de varias especies de escorpiones, estableciendo como variables las distribuciones de las k-tuplas con otros métodos multivariados como el Análisis Factorial Múltiple con el objetivo de encontrar similitudes entre grupos de especies.

Bibliografía

Almeida, J. S., Carrico, J. A., Maretzek, A., Noble, P. A. & Fletcher, M. (2001), 'Analysis of genomic sequences by Chaos Game Representation', *Bioinformatics* **17**(5), 429–437. https://doi.org/10.1093/bioinformatics/17.5.429.

Alvarenga, É. R., Mendes, T. M., Magalhães, B. F., Siqueira, F. F., Dantas, A. E., Barroca, T. M., Horta, C. C. & Kalapothakis, E. (2012), 'Transcriptome analysis of the Tityus serrulatus scorpion venom gland', *Open Journal of Genetics* (2), 210–220. Scientific Research Publishing. http://dx.doi.org/10.4236/ojgen.2012.24027.

Amiri, S. & Dinov, I. D. (2017), 'msktuple: An integrated R library for alignment-free multiple sequence k-tuple analysis', *Chemometrics and Intelligent Laboratory Systems* **168**, 84–88. https://doi.org/10.1016/j.chemolab.2017.07.012.

Aria, M. & Cuccurullo, C. (2017), 'bibliometrix: An R-tool for comprehensive science mapping analysis', *Journal of informetrics* **11**(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007.

Baena, D. D. S. R., Santos, D. J. C. R. & Ruiz, D. J. S. A. (2006), Análisis de datos de Expresión Genética mediante técnicas de Biclustering, PhD thesis, PhD thesis, Dto. Lenguajes y Sistemas Informáticos, Universidad de Sevilla.

Comin, M., Leoni, A. & Schimd, M. (2015), 'Clustering of reads with alignment-free measures and quality values', *Algorithms for Molecular Biology* **10**(1), 1–10. https://doi.org/10.1186/s13015-014-0029-x.

Cooper, G. M. & Hausman, R. E. (2010), *La célula*, 5 edn, Marbán. Edición en español de: *The Cell: a Molecular Approach, 5th edition.*

Dawkins, R. (1982), *The Extended Phenotype [El fenotipo extendido]*, Vol. 8, Oxford University Press.

Escobar, C. A. M., Murillo, L. V. R. & Soto, J. F. (2011), 'Tecnologías bioinformáticas para el análisis de secuencias de ADN', *Scientia et technica* **3**(49), 116–121.

Han, J., Pei, J. & Kamber, M. (2011), *Data mining: concepts and techniques*, Elsevier.

Karp, G. (2005), *Biologia celular e molecular*, Editora Manole Ltda.

Kassambara, A. (2017), *Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra*, Vol. 2, STHDA.

Kazemi-Lomedasht, F., Khalaj, V., Bagheri, K. P., Behdani, M. & Shahbazzadeh, D. (2017), 'The first report on transcriptome analysis of the venom gland of Iranian scorpion, Hemiscorpius lepturus', *Toxicon* **125**, 123–130. https://doi.org/10.1016/j.toxicon.2016.11.261.

Konishi, T., Matsukuma, S., Fuji, H., Nakamura, D., Satou, N. & Okano, K. (2019), 'Principal component analysis applied directly to sequence matrix', *Scientific reports* **9**(1), 1–13. https://doi.org/10.1038/s41598-019-55253-0.

Lahoz-Beltrá, R. (2010), *Bioinformática: Simulación, vida artificial e inteligencia artificial*, Ediciones Díaz de Santos.

Lê, S., Josse, J. & Husson, F. (2008), 'FactoMineR: an R package for multivariate analysis', *Journal of statistical software* **25**(1), 1–18. https://doi.org/10.18637/jss.v025.i01.

Lebart, L., Morineau, A. & Piron, M. (1995), *Statistique exploratoire multidimensionnelle*, Vol. 3, Dunod Paris.

Li, M., Badger, J. H., Chen, X., Kwong, S., Kearney, P. & Zhang, H. (2001), 'An information-based sequence distance and its application to whole mitochondrial genome phylogeny', *Bioinformatics* **17**(2), 149–154. https://doi.org/10.1093/bioinformatics/17.2.149.

Ma, Y., Zhao, R., He, Y., Li, S., Liu, J., Wu, Y., Cao, Z. & Li, W. (2009), 'Transcriptome analysis of the venom gland of the scorpion Scorpiops jendeki: implication for the evolution of the scorpion venom arsenal', *BMC genomics* **10**(1), 290. https://doi.org/10.1186/1471-2164-10-290.

Ochoa Muñoz, A. F. (2018), Análisis de Correspondencias Múltiples en presencia de datos faltantes: el principio de datos disponibles del algoritmo NIPALS (ACMpdd), Master's thesis, Universidad del Valle.

OMS (2007), 'La OMS prevé aumentar el acceso al tratamiento para las víctimas de la rabia o de mordeduras de serpiente'.
URL: *https://www.who.int/mediacentre/news/notes/2007/np01/es/*

Pardo, C. E. & Del Campo, P. C. (2007), 'Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete FactoClass', *Revista colombiana de estadística* **30**(2), 231–245.

Polis, G. A. (1990), *The biology of scorpions*, number 595.46 B5.

Rodriguez, M. & Proyecto, I. (2015), 'Comparación de métricas de distancia en el algoritmo K-Vecinos Más cercanos para el problema de Reconocimiento Automático de Dígitos Manuscritos', *Pontificia Universidad Católica de Valparaíso, Facultad de Ingeniería, Escuela de Ingeniería Informática* .

Schwartz, E. F., Diego-Garcia, E., de la Vega, R. C. R. & Possani, L. D. (2007), 'Transcriptome analysis of the venom gland of the Mexican scorpion Hadrurus gertschi (Arachnida: Scorpiones)', *BMC genomics* **8**(1), 119. https://doi.org/10.1186/1471-2164-8-119.

Starr, C. & Taggart, R. (2008), *Biología la unidad y la diversidad de la vida*, Vol. 8, Instituto Politécnico Nacional.

Villamil, M. L. A. (2016), 'Veneno de escorpión colombiano podría eliminar células cancerígenas', *Un periódico* (198), 12–13. Universidad Nacional de Colombia.

Vinga, S. & Almeida, J. (2003), 'Alignment-free sequence comparison—a review', *Bioinformatics* **19**(4), 513–523. https://doi.org/10.1093/bioinformatics/btg005.