# Relación evolutiva de los intrones con base en el Modelo de Pérdida de ADN para las familias de neuropéptidos LWamida, APGWamida, Hormona Concentradora de Pigmentos Rojos, Hormona Adipocinética, Corazonina y Hormona Liberadora de Gonadotropina

María Alejandra Reyes Tarazona

Trabajo de Grado para Optar al Título de Bióloga

Director

Francisco José Martínez Pérez

Doctor en Ciencias

Codirector

Laura Rebeca Jiménez Gutiérrez

Doctora en Ciencias

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Biología

Bucaramanga

2022

**Dedicatoria**

*A mi madre Laura Consuelo Tarazona Serrano: excelente persona, calidad de mujer, quien con amor me enseñó a creer en mí, fue su alegría y apoyo incondicional; lo que me llevó a perseverar en momentos difíciles. Tu más que nadie te mereces este logro…*

*¡Esto es por ti, que eres luz en mi vida!*

**Agradecimientos**

Agradezco el desarrollo de esta pasantía al ser parte del proyecto titulado: “Aplicación del modelo evolutivo de pérdida de ADN en genes neuroendocrinos con fines en ciencia básica y biomédica”. A los profesores de la escuela de Biología de la Universidad Industrial de Santander, por brindarme a lo largo de la carrera las bases y los fundamentos para formarme como profesional en Biología.

Al Dr. Francisco José Martínez Pérez por su vocación y amor por la ciencia, por ser guía en este proyecto, por sus palabras de aliento, por enseñarme con paciencia y sobre todo por creer en mí.

Al Dr. Carlos Jaime Barrios Hernández por permitirme ser parte de su grupo de trabajo. Al Grupo de Investigación Computo Avanzado y a Gran Escala (CAGE).

A mi padre Reynaldo Reyes quien es mi orgullo. Tu forjaste el carácter y la determinación de la mujer que soy, fuiste la inspiración que me llevó a emprender en este camino.

A mi hermano Jhon Alexander Reyes. Porque no se puede pintar la vida de colores, sin antes tener un bosquejo… Gracias por dibujar los cimientos de mi ser.

A Cristian Enrique Cadena Caballero por su dedicación y guía en el desarrollo de este trabajo.

A amigos y familiares que fueron mi fuerza y me brindaron valiosos momentos en cada paso.

A todos aquellos que aman la ciencia y su vivir está en el compartir el conocimiento.

**Tabla de Contenido**

**Pág**.

## Lista de tablas

**Pág.**

## Lista de Figuras

**Pág.**

## Lista de Apéndices

**Apéndice A**. Búsqueda e identificación de secuencias que presenten gen mediante código al GenBank.

**Apéndice B.** Caracterización de los intrones y sitios proto-empalme.

**Apéndice C.** Alineamiento de las familias LW, APGW y APGW-AKH.

**Apéndice D.** Alineamiento de las familias AKH, APGW y APGW-AKH.

**Apéndice E.** Alineamiento de la familia CRZ.

**Apéndice F.** Alineamiento de la familia ACP.

**Apéndice G.** Alineamiento de la familia RPCH.

**Apéndice H**. Alineamiento de la familia GnRH.

**Apéndice I.** Secuencias Morado AKH_RPCH.

**Apéndice J.** Base de datos de genes neuropéptidicos.

**Apéndice K.** Alineamientos Gen_ARNm_Amino.

*Nota:* Los apéndices están adjuntos y puede visualizarlos en la base de datos de la biblioteca UIS.

## Resumen

**Título:** Relación evolutiva de los intrones con base en el Modelo de Pérdida de ADN para las familias de neuropéptidos LWamida, APGWamida, Hormona Concentradora de Pigmentos Rojos, Hormona Adipocinética, Corazonina y Hormona Liberadora de Gonadotropina *

**Autor:** María Alejandra Reyes Tarazona**

**Palabras Clave:** Intrones, Evolución de intrones, Fase intrón, Sitios Proto-empalme, Evolución de neuropéptidos, DNA-LM.

**Descripción:** Los mecanismos celulares que regulan la pérdida y ganancia de intrones son ampliamente estudiados, ya que dan indicios de cómo se generó la diversidad de la vida. Así, diversos modelos evolutivos han planteado el origen de genes y sus posibles causas. Ejemplo de esto, son las hipótesis de intrones o los supuestos formulados para la evolución de la Hormona Adipocinética (AKH), Corazonina (CRZ), Hormona Liberadora de Gonadotropina (GnRH) y precursores híbridos como la Hormona Adipocinética/Corazonina (ACP). Sin embargo, solo el Modelo de Pérdida de ADN (DNA-LM), describe el origen de genes neuroendocrinos a partir de perdida de nucleótidos y movimiento de intrones. Aquí determinamos *in silico* con los principios del DNA-LM, las posiciones, fases intrónicas y sitios de Proto-empalme (*Protosplice)* de los genes de las familias de neuropéptidos antes mencionadas junto con la LWamida (LW), APGWamida (APGW), Hormona Concentradora de Pigmentos Rojos (RPCH) y el precursor híbrido APGW/AKH, para establecer sus relaciones filogenéticas por medio de la conservación de intrones ortólogos y su evolución por pérdida y/o ganancia de intrones. Los resultados presentaron un intrón ortólogo en casi todas las secuencias, para la LW se cree que lo perdió en su desarrollo. Por otro lado, se encontró un patrón de distribución de las fases, los intrones fase 1 y 2 se ubican mayormente en el extremo 5', donde se determinaron nuevos intrones, mientras los de fase 0 que fueron predominantes, se asentaban en el extremo 3', en el intrón conservado; lo que sugirió que los nuevos intrones no se integran aleatoriamente. Finalmente, se logró establecer la relación ancestral de estas familias neuropéptidicas por medio del intrón ortólogo. Los intrones recientes y los sitios remanentes de proto-empalme en ORF que no poseen intrones, fueron evidencia de la evolución independiente de cada linaje, lo que contribuyó a corroborar el DNA-LM.

---

* Trabajo de Grado

** Facultad de Ciencias. Escuela de Biología.
Director: Francisco José Martínez Pérez. Doctor en Ciencias.
Co-directora: Laura Rebeca Jiménez Gutiérrez. Doctora en Ciencias.

## Abstract

**Title:** Evolutionary relationship of introns based on the DNA Loss Model for the neuropeptide families LWamide, APGWamide, Red Pigment Concentrating Hormone, Adipokinetic Hormone, Corazonin and Gonadotropin Releasing Hormone [*]

**Author(s):** María Alejandra Reyes Tarazona[**]

**Key Words:** Introns, Intron evolution, Intronic phase, Protosplice sites, Neuropeptide evolution, DNA-LM.

**Description:** The cellular mechanisms that regulate the loss and gain of introns are widely studied, since they provide clues about how the diversity of life. Thus, various evolutionary models have been proposed on the origin of genes and their possible causes. An example of the above are the different hypotheses of introns or the assumptions made for the evolution of Adipokinetic Hormone (AKH), Corazonin (CRZ), Gonadotropin-Releasing Hormone (GnRH) and hybrid precursors such as Adipokinetic Hormone/Corazonin (ACP). However, only the DNA Loss Model (DNA-LM) describes the origin of neuroendocrine genes from nucleotide loss and intron movement. By *in silico* analysis with the principles of DNA-LM, we determined the positions, intronic phases and protosplices sites of the genes encoding the neuropeptide families together with LWamide (LW), APGWamide (APGW), Red Pigment Concentrating Hormone (RPCH) and the hybrid precursor APGW/AKH, to establish their phylogenetic relationships through the conservation of orthologous introns and their evolution by loss and/or gain of introns. The results presented an orthologous intron in almost all the sequences, for LW it was considered that it was lost in its development. On the other hand, a distribution pattern of the phases was displayed; phase 1 and 2 introns were located mostly at the 5' end, where new introns were observed. Meanwhile the predominant phase 0 introns were found at the 3’ end in the conserved intron, which suggested that new introns do not integrate randomly. Finally, it is possible to establish the ancestral relationship of these neuropeptide families through the orthologous intron. The recent introns and the remnant sites of proto splicing in ORFs that did not possess introns, was evidence of the independent evolution of each lineage, which contributed to the DNA-LM corroboration.

---

[*] Bachelor thesis

[**] Science Faculty. School of Biology.
Adviser: PhD. Francisco José Martínez Pérez.
Adviser: PhD. Laura Rebeca Jiménez Gutiérrez.

## Introducción

Los genes en las células eucariotas tienen regiones llamadas exones que codifican para proteínas y segmentos intermedios denominados intrones (Gilbert, 1978). Estos últimos son secuencias que en un principio se creía que eran “basura” al ser descartadas por la célula misma. Sin embargo, diferentes estudios le han atribuido funciones como el almacenamiento de elementos reguladores (Rogozin et al., 2012), eficiencia en la formación de proteínas (Jo y Choi, 2015; Lim et al., 2018), potenciadores en la recombinación homóloga (Fedorova et al., 2003); incluso algunos autores mencionan una probable actividad ancestral, la cual, actualmente ya no es vigente (Chorev y Carmel, 2012).

A nivel evolutivo los intrones juegan un papel realmente importante, ya que se descubrió que la posición de un intrón dentro del gen, se heredaba (Rogozin et al., 2012). En este sentido, si un intrón concuerda en la ubicación en un codón y/o regiones no traducidas del ARNm en grupos aislados y cercanos filogenéticamente, refleja un estado de conservación que se mantiene y se hereda, a estos, se les conoce como intrones ortólogos (Rogozin et al., 2012). Por otro lado, si un intrón se encuentra en una posición solamente para una especie o especies cercanas filogenéticamente se entienden como intrones que no se heredaron de un gen ancestral, sino que probablemente se han integrado recientemente al genoma (Rogozin et al., 2012). De esta manera la identificación de los intrones en los genomas es de gran consideración, para lo cual, existen mecanismos de definición intrón y exón, estos han sido descritos a partir del proceso químico de reacción y de reconocimiento por parte del espliceosoma (Jiménez-García et al., 2004). Una de las características intrínsecas de definición de un intrón es la secuencia canónica |GT…AG|, esta se

encuentra conservada en un 98% de los intrones en todo tipo de genomas eucariontes, también existen otras secuencias no canónicas, pero estas se encuentran en porcentajes bajos (Vinogradov, 2006; Poverennaya y Roytberg, 2020).

Los intrones también pueden ser clasificados según su ubicación, para esto es necesario precisar la posición dentro del codón, en el cual, se encuentra integrado el intrón. Según lo anterior, si el intrón se encuentra entre el primer y segundo nucleótido es fase 1, si está entre el segundo y tercer nucleótido es fase 2 y, si está presente entre dos codones es fase 0 (Long et al.,1998). Las distinciones de las fases han sido estudiadas en correlación con los límites exón/intrón y con la conservación en la localización de intrones (Long et al.,1998; Nguyen et al., 2006; Poverennaya et al.,2017).

El descubrimiento de los intrones en 1977 (Chow et al., 1977), trajo un gran número de cuestionamientos, uno de los que más se destacan es con base a su origen, el cual, ha generado un gran debate en el último siglo (Gilbert, 1978; Logsdon, 1998; Koonin, 2006; Rogozin et al., 2012). Inicialmente, se planteó la hipótesis de los “intrones tempranos”, la cual, postula que las células eucariotas heredaron sus intrones de las procariotas y que la diversidad genómica entre los eucariontes se debía a la pérdida de intrones en la historia individual de cada linaje (Gilbert, 1987; Giovannoni, 2014); en el caso de las procariotas actuales la “racionalización del genoma” (Giovannoni, 2014), fue lo que generó la pérdida de sus intrones primordiales y del espliceosoma (Rogozin et al., 2012).

Posteriormente, surgió la hipótesis de “intrones tardíos”, que estableció la integración de “nuevos intrones” en posiciones específicas del genoma, llamados sitios de proto-empalme (*protosplice*) o límites exón/intrón (Dibb y Newman, 1989; Gilbert, 1993; Sverdlov et al., 2004). Esta hipótesis propuso que las células procariotas nunca presentaron intrones, siendo una novedad

eucarionte, en la cual, con el transcurso del tiempo “nuevos intrones” se integraron, ocasionando las diferencias entre sus genes homólogos (Logsdon, 1998; Koonin, 2006; Stoltzfus, 1997). A estas conjeturas, se le realizaron adaptaciones, por ejemplo, la teoría de los “primeros intrones”, que indica que los intrones y exones no se originaron simultáneamente, sino que los intrones precedieron a los exones en el “mundo del ARN”, es decir, los ARNm eran totalmente intrónicos y comenzaron a llamarse intrones al surgir los límites exónicos (Forsdyke, 2013).

Los postulados de evolución de intrones sirvieron como uno de los fundamentos para generar el modelo de origen de genes neuroendocrinos “Pérdida de ADN (DNA-LM)” (Martínez-Pérez et al., 2007).

En contexto con el modelo es necesario entender la estructura y formación de neuropéptidos. La biosíntesis de neuropéptidos ocurre de la misma forma que la de proteínas de secreción, después de la transcripción del gen que codifica para el precursor neuropéptidico (conformado por: péptido señal - péptido activo - péptido relacionado), se producirá una o más copias de pre-ARNm (Garcia-Lopez et al., 2002), seguidamente se dará inicio al proceso de maduración del ARNm o corte y empalme nombrado en inglés *splicing*. El espliceosoma que es una ribonucleoproteína, reconocerá señales claves en el exón y en el intrón, las cuales, son complementarias a sus subunidades e indicarán en donde escindir y que empalmar (Jiménez et al., 2004). De esta forma, los intrones son desechados y los exones son unidos por la generación del enlace fosfodiéster. Finalmente, el ARNm ya maduro será traducido por los ribosomas adheridos al retículo endoplásmico rugoso y formará los neuropéptidos vía aparato de Golgi (Garcia-Lopez et al., 2002).

Partiendo de lo anterior, en el 2002 se realizó un estudio de la relación entre secuencias de aminoácidos de tres familias de neuropéptidos: la AKH de insectos, la RPCH de crustáceos y la

APGWamida de moluscos; los autores concluyeron que a partir de un posible gen ancestral se dio origen a estas familias (Martínez-Pérez et al., 2002). Los investigadores mencionaron que este gen podría estar conformado por tres exones y dos intrones, que, con el tiempo, debido al ingreso y deleción de intrones y a la pérdida de nucleótidos y fragmentos de ADN en el primer exón, dio origen a la APGW de moluscos. Posteriormente, este proceso se repitió en el segundo exón, lo que generó la formación de la RPCH de crustáceos, la cual, se conservó hasta cierto grado en las AKH de insectos (Martínez-Pérez et al., 2002). Para el 2007, se añade la familia LW de hidras a la investigación, ya que se observó que el triptófano estructural (W) generaba la actividad biológica en el péptido activo del precursor (Cadena-Caballero, 2020) y que, a partir de los genes de esta y por medio de duplicación y pérdida de ADN se generó a la APGW de moluscos. Lo anterior, dio paso a la postulación del DNA-LM (Martínez-Pérez et al., 2007).

El DNA-LM, propone que la evolución de genes de neuropéptidos ocurrió por la duplicación de un gen ancestral que codificó para un precursor neuropeptídico híbrido con distinto número de copias y, que debido al movimiento de intrones y a la pérdida de codones se generaron nuevos dominios a nivel molecular (Martínez-Pérez et al., 2002; Martínez-Pérez et al., 2007; Cadena-Caballero, 2020). El DNA-LM se planteó para familias neuropéptidicas que contienen aminoácidos homólogos conservados evolutivamente en los extremos carboxilos, como lo son LW de hidras, APGW de moluscos, RPCH de crustáceos, AKH de insectos y sus péptidos híbridos. Estos últimos, son secuencias intermedias que presentan características similares entre dos o más neuropéptidos, como lo son la Hormona Adipocinética/Corazonina-péptido relacionado (ACP) de insectos y crustáceos y las teóricas APGW/RPCH/AKH y APGW/RPCH de moluscos (Martínez-Pérez et al., 2002; Martínez-Pérez et al., 2007; Hansen et al., 2010). Posteriormente, a partir de los proyectos genómicos de Secuenciación de Nueva Generación (NGS), se demostró que las familias

de neuropéptidos CRZ en especies de invertebrados y GnRH en cordados; tienen relación evolutiva con las familias de genes APGW de moluscos, RPCH de crustáceos y AKH de insectos (Cadena-Caballero, 2020).

Los supuestos anteriores han evaluado la evolución por intrones desde un punto general en genomas completos (Lynch., 2002; Csuros et al., 2011; Sêton et al., 2016; Mukhopadhyay y Hausner., 2021) o en particular, en genes específicos que codifican para proteínas muy conservadas (Roesner et al., 2005; Parenteau et al., 2011). Sin embargo, estos no han considerado la posición de los intrones, sitios de proto-empalme y fases intrón en grupos aislados y/o cercanos filogenéticamente para las familias antes descritas. El DNA-LM es un modelo evolutivo basado en la pérdida de regiones de ADN, el cual, propone que la influencia y el movimiento de intrones ha generado la variación en genes neuroendocrinos. Por otro lado, los estudios de evolución por intrones buscan encontrar concordancia entre los árboles filogenéticos de genes específicos y árboles de especies (Creer, 2007). Previamente el modelo debió establecer las relaciones filogenéticas entre estas familias. Sin embargo, aún falta corroborarlo con el análisis de intrones, razón por la cual se desarrolló este proyecto. De esta manera, el presente estudio con base en el DNA-LM, de las familias LW de hidras, APGW de moluscos, RPCH de crustáceos, AKHs de insectos, CRZ de invertebrados, GnRH de cordados y sus precursores híbridos determinó: posiciones, fases intrónicas y la homología de los sitios en los límites exón/intrón, para establecer su relación por medio de la conservación en la posición de intrones ortólogos y su evolución en especies de invertebrados y cordados, a partir de la ganancia o pérdida de intrones.

# 1. Objetivos

## 1.1 Objetivo de la pasantía

Establecer la relación evolutiva de las familias de neuropéptidos LW, APGW, RPCH, AKH, CRZ, GnRH y precursores híbridos con el DNA-LM por medio de pérdida y/o ganancia de intrones.

## 1.2 Objetivo General

Determinar las posiciones intrónicas y límites exónicos de los genes que codifican a las familias de neuropéptidos: LW, APGW, RPCH, AKH, CRZ, GnRH y precursores híbridos para contribuir al DNA-LM, al establecer la influencia de la pérdida de ADN en la evolución.

## 1.3 Objetivos Específicos

Generar una base de datos de los genes LW, APGW, RPCH, AKH, CRZ, GnRH y precursores híbridos reportados en el GenBank.

Caracterizar la organización de los intrones y exones de los genes LW, APGW, RPCH, AKH, CRZ, GnRH y precursores híbridos reportados en el GenBank.

Correlacionar las posiciones de los intrones y límites exónicos de los genes que codifican a las familias: LW, APGW, RPCH, AKH, CRZ, GnRH y precursores híbridos con el DNA-LM.

# 2. Competencias adquiridas

## 2.1 Competencias cognitivas

Identifica genes para precursores de neuropéptidos híbridos con bioinformática.

Comprende los términos intrón y exón para identificar en genes de neuropéptidos.

Caracteriza intrones y exones para establecer su organización.

Correlaciona posiciones de intrones y límites exónicos para la evolución de intrones de familias de neuropéptidicas.

Relaciona los intrones de las familias de neuropéptidos para la evolución de genes con el DNA-LM.

## 2.2 Competencias procedimentales

Construye una base de datos de genes de neuropéptidos sustentada en intrones y exones para las familias: LW, APGW, RPCH, AKH, CRZ, GnRH y precursores híbridos.

Estructura los genes al reconocer los límites intrón/exón en genes de neuropéptidos.

## 2.3 Competencias actitudinales

Comprende la importancia del trabajo en equipo, respetando las diferentes opiniones para el buen uso de ellas.

Atiende a las críticas constructivas para mejorar su desarrollo como profesional y futura bióloga.

Genera espacios de respeto y tolerancia para con sus compañeros y directores.

# 3. Metodología

## 3.1 Generación de la base de datos de los genes LW, APGW, RPCH, AKH, CRZ, GnRH y precursores híbridos reportados en el GenBank

Las secuencias de los genes que codifican para precursores de las familias LW de hidras, APGW de moluscos, RPCH de crustáceos, AKH de insectos, CRZ de invertebrados y GnRH de cordados fue generada a partir de la base de datos de precursores de neuropéptidos del Laboratorio de Genómica Celular y Aplicada (LGCA) del Grupo de Cálculo Avanzado y a Gran Escala (CAGE) de la Universidad Industrial de Santander, esta contiene las secuencias de aminoácidos para cada precursor y su código de acceso; a partir de este código se realizó una búsqueda en la base de datos pública GenBank del National Center for Biotechnology Information (NCBI) (Sayers et al., 2020), empleando solo aquellas secuencias que presentaban el gen. Adicionalmente se utilizó un parámetro de descarte con las secuencias que se reportaron en el GenBank como "Versiones obsoletas" y se renovó la información de secuencias que ya habían sido actualizadas en el GenBank.

En el caso de las secuencias que se identificaron que presentaban gen, fueron descargadas en formato FASTA del GenBank y se tabuló la información de estos en un documento de Microsoft Excel (2016). Las secuencias de los precursores híbridos, se obtuvieron con un alineamiento tipo BLAST (*Basic Local Alignment Search Tool*), respecto a los péptidos teóricos iniciales del DNA-LM mediante las secuencias de aminoácidos de los precursores en el GenBank, el gen se identificó como se indicó previamente (Altschul et al., 1990; Martínez-Pérez et al., 2007).

### 3.2 Caracterización de la organización de los intrones y exones de los genes LW, APGW, RPCH, AKH, CRZ, GnRH y precursores híbridos reportados en el GenBank

Para identificar la forma en la que estaban dispuestos los exones y los intrones en el gen, se descargó el ARNm de cada neuropéptido en formato FASTA, este procedimiento se llevó a cabo de la misma forma descrita en el punto 3.1. Con la secuencia del gen y del ARNm, se procedió a realizar un alineamiento con el programa CLUSTAL W (Larkin et al., 2007) y los parámetros de DNA-LM (Martínez-Pérez et al., 2007), a partir de allí se generó un archivo de texto plano. En la lectura del alineamiento se logró distinguir los exones de los intrones, esto se realizó teniendo en cuenta el mecanismo de definición Exón/Intrón, en el cual, se precisó los sitios de unión entre estos, que a su vez son las señales de reconocimiento de corte y empalme por el espliceosoma, previamente descritos por otros autores (Dibb y Newman, 1989; Vinogradov 2006; Poverennaya y Roytberg 2020). Posteriormente, se tomó la secuencia del ARNm y se tradujo en la página *Translate tool-Expasy* (https://web.expasy.org/translate/), para reconocer el aminoácido de cada codón y se colocó manualmente en el documento del alineamiento del gen con el ARNm.

Se realizó para cada secuencia su clasificación respecto a las fases intrónicas dependiendo de la posición del codón en donde se integraba el intrón. Es decir, si el intrón se insertó entre el primer y segundo nucleótido de un codón, sería Fase 1, si estaba presente entre el segundo y tercer nucleótido de un codón, se llamaría Fase 2 y si se introdujo entre dos codones, se denominaría Fase 0 (Long et al.,1998). Las regiones que corresponden a los límites exónicos e intrónicos de los genes de las secuencias identificadas fueron tabuladas en un documento de Microsoft Excel (2016), donde se incluyó: la familia de neuropéptido, la especie, el código de acceso del GenBank, la fase del intrón, la cantidad de intrones dentro del gen y el aminoácido en donde estaba presente el intrón o aminoácidos en caso de la fase 0. Para este estudio no se tuvieron en cuenta las Regiones

No Traducidas (UTRs), solo se analizó la organización de los intrones y exones dentro del Marco Abierto de Lectura (ORF).

### 3.3 Correlación de las posiciones de los intrones y límites exónicos de los genes que codifican a las familias: LW, APGW, RPCH, AKH, CRZ, GnRH y precursores híbridos con el DNA-LM

A fin de correlacionar las posiciones intrónicas y los sitios de proto-empalme de estas familias con el DNA-LM, se empleó un alineamiento con el programa CLUSTAL W por cada una de las familias de las secuencias aminoacídicas analizadas en este estudio. Aquellas familias neuropeptídicas que presentaron solo una secuencia, se alinearon con las secuencias de las familias más similares, seguido de esto, cada alineamiento fue editado por el programa GeneDoc (Wenger y Mathonet, 2002). Posteriormente con el programa Word, se resaltó la posición de los intrones ya determinados en el paso 3.2, con un color en específico dependiendo de la fase (0: amarillo, 1: azul aguamarina y 2: verde). Para precisar la posición de los intrones en las secuencias, se contó como ortólogos aquellos que se encontraron en un rango entre 1 a 5 aminoácidos, en relación con los intrones del resto de secuencias por familia y se tuvo en cuenta el deslizamiento de intrones descrito por Tarrio et al (2008).

En función de la posición de los intrones mostrados en el alineamiento, a modo de representarlos esquemáticamente se distinguieron los intrones ortólogos de los nuevos intrones a partir del porcentaje de secuencias por familia que los ubicaban en la misma posición, en este sentido, se empleó los criterios establecidos por el grupo CAGE, si el intrón se localizaba en el mismo punto para más del 50% de las secuencias por familia, se consideró como ortólogo y si se

encontraba en menos del 50%, se categorizó como reciente, estos últimos representan una forma de pérdida o ganancia de nucleótidos.

# 4. Resultados

## 4.1 Base de datos de los genes de LW, APGW, RPCH, AKH, CRZ, GnRH y sus precursores híbridos

De las 877 secuencias que conforman la base de datos de precursores de neuropéptidos del LGCA adjunto al grupo CAGE, solo 252 presentaron su gen correspondiente en el GenBank. En este sentido, para cada familia neuropeptídica, fueron pocas las secuencias confirmadas de acuerdo al número total de secuencias reportadas: LW de hidras (27/37), APGW de moluscos (1/12), RPCH de crustáceos (3/18), AKH1 (79/172), AKH2 (3/15), AKH3 (23/73), AKH4 (1/4), CRZ de invertebrados (10/62), GnRH de cordados (76/436) y del precursor híbrido ACP (29/48). La cantidad de especies resultantes de la filtración de la base de datos inicial fue de 178, donde en su mayoría estaban representadas por las familias AKH de insectos y GnRH de cordados con 97 y 41 respectivamente, seguidas de 17 por parte de la ACP y 10 de la LW de hidras; el resto de las familias quedaron con menos de 10 especies.

Por otro lado, en total 32 especies presentaron empalme alternativo, 6 fueron actualizadas por los autores originales en la base de datos del GenBank, 9 se descartaron por ser versiones obsoletas y se encontró 14 especies que tenían más de una secuencia que daban para genes diferentes. En este sentido, se observó que casi todas las especies pertenecientes a las AKHs de insectos constaban de un solo ARNm, mientras que en muchas de las especies de las familias GnRH de cordados y ACP poseían más de una variante o más de un gen.

El alineamiento tipo BLAST de los precursores híbridos propuestos en el DNA-LM en 2007 mostró homología con una secuencia perteneciente a la especie *Brachionus plicatilis.* Esta secuencia presenta similaridad en su péptido activo con el de las AKHs de insectos, además de poseer dos copias de APGW de moluscos. Tales características que comparte con estas dos familias le confieren el nombre de precursor híbrido APGW/AKH (VP- APGW/AKH = HyPro-BpHYR1_030954). Esta secuencia fue agregada a la base de datos de precursores neuropeptídicos quedando con 878 secuencias y a la base de datos de genes con un total de 253 secuencias. La depuración de la base de datos de precursores neuropeptídicos con el parámetro de presencia del gen, generó un bajo número de secuencias para las familias APGW de moluscos y RPCH de crustáceos. Esto podría catalogarse con un sesgo hacia las familias AKH de insectos y GnRH de cordados, las cuales tuvieron una mayor representación en este estudio (Apéndice A y J).

### 4.2 Organización de los intrones y exones en los genes LW, APGW, RPCH, AKH, CRZ, GnRH y sus precursores híbridos.

De las 253 secuencias analizadas pertenecientes a 179 especies, el 48.62% presentaban dos intrones, con una mayor presencia en las secuencias de GnRH de cordados. La familia AKH de insectos fue la que presentó más secuencias con un solo intrón y la familia LW de hidras mostró más secuencias con un solo exón. La secuencia del precursor híbrido APGW/AKH poseía 3 exones y 2 intrones. Solamente 2 secuencias de GnRH de cordados poseían hasta 5 intrones, mientras que el resto de las secuencias analizadas poseían 3 intrones o menos (Tabla 1. Apéndice B).

**Tabla 1**

*Número de intrones presentes en cada familia neuropeptídica.*

| Neuropéptidos | No. Especies | Cantidad de intrones por secuencia | | | | | | No. de secuencias totales |
|---|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** | **4** | **5** | |
| **APGW/AKH** | 1 | 0 | 0 | 1 | 0 | 0 | 0 | **1** |
| **LW** | 10 | 20 | 7 | 0 | 0 | 0 | 0 | **27** |
| **APGW** | 1 | 0 | 1 | 0 | 0 | 0 | 0 | **1** |
| **RPCH** | 3 | 0 | 3 | 0 | 0 | 0 | 0 | **3** |
| **AKH** | 97 | 5 | 70 | 29 | 2 | 0 | 0 | **106** |
| **CRZ** | 9 | 0 | 9 | 1 | 0 | 0 | 0 | **10** |
| **GnRH** | 41 | 1 | 0 | 69 | 4 | 0 | 2 | **76** |
| **ACP** | 17 | 0 | 1 | 23 | 5 | 0 | 0 | **29** |
| **Total** | **179** | **26** | **91** | **123** | **11** | **0** | **2** | **253** |
| **%** | | **10.28%** | **35.97%** | **48.62%** | **4.35%** | **0.00%** | **0.79%** | **100.00%** |

*Nota:* Se muestra el número de especies y secuencias por familia. Además de la cantidad y el porcentaje de secuencias que presentaban un número determinado de intrones.

La mayoría de las secuencias analizadas contenían intrones de fase 0 (79.47%). Dado el sesgo en la cantidad de secuencias superior por parte de las AKHs de insectos y de las GnRHs de cordados en comparación con las demás familias, estas son las familias con más intrones en total. Con excepción de la familia LW de hidras que todos sus intrones fueron fase 2, la mayoría de las familias neuropeptídicas presentó principalmente intrones en fase 0. La secuencia APGW/AKH presentó 2 intrones, uno fase 0 y el otro en fase 1. La familia ACP fue la que presentó mayor número de intrones en fase 2 y a su vez la familia GnRH de cordados fue mayoritaria en intrones fase 1 (Tabla 2 y Apéndice B).

**Tabla 2**

*Porcentaje de las fases intrónicas para cada neuropéptido*

| Neuropéptidos | Fase Intrón | | | Total Intrones |
|---|---|---|---|---|
| | **0** | **1** | **2** | |
| APGW/AKH | 1 | 1 | 0 | 2 |
| LW | 0 | 0 | 7 | 7 |
| APGW | 1 | 0 | 0 | 1 |
| RPCH | 3 | 0 | 0 | 3 |
| AKH | 116 | 2 | 16 | 134 |
| CRZ | 7 | 1 | 3 | 11 |
| GnRH | 139 | 17 | 4 | 160 |
| ACP | 35 | 0 | 27 | 62 |
| **Total** | **302** | **21** | **57** | **380** |
| **%** | **79.47%** | **5.53%** | **15.00%** | **100.00%** |

*Nota:* La tabla muestra la cantidad de intrones que se encontraron por familia y el número de intrones por fase.

En cuanto a los límites de los exones con los intrones, el sitio de proto-empalme Exón | Intrón que más fue conservado entre las secuencias fue G|GT (74.14%) y para el caso de los porcentajes de los límites Intrón | Exón estos fueron más heterogéneos, sobresaliendo solamente el límite: AG|A (44.59%), lo que denota un claro porcentaje mayor de purinas entre estos límites. (Tabla 3, Apéndice B).

**Tabla 3**

*Sitios de proto-empalme para las familias neuropeptídicas*

| Neuropéptidos | Nucleótidos de los Proto-empalme | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Exón \| Intrón | | | | Intrón \| Exón | | | |
| | A\|GT | C\|GT | G\|GT | T\|GT | AG\|A | AG\|C | AG\|G | AG\|T |
| **APGW/AKH** | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| LW | 1 | 0 | 5 | 1 | 5 | 0 | 2 | 0 |
| APGW | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| RPCH | 0 | 0 | 3 | 0 | 1 | 0 | 2 | 0 |
| AKH | 24 | 1 | 98 | 10 | 57 | 35 | 27 | 14 |
| CRZ | 2 | 0 | 8 | 1 | 0 | 8 | 0 | 3 |
| GnRH | 9 | 2 | 120 | 29 | 82 | 17 | 42 | 19 |
| ACP | 12 | 2 | 45 | 3 | 23 | 16 | 22 | 1 |
| **Total** | **49** | **5** | **281** | **44** | **169** | **77** | **96** | **37** |
| **%** | **12.93%** | **1.32%** | **74.14%** | **11.61%** | **44.59%** | **20.32%** | **25.33%** | **9.76%** |

*Nota:* Los sitios de Proto-empalme Exón | Intrón hacen referencia al último nucleótido del exón donador y a los dos primeros nucleótidos del intrón. Los sitios de Proto-empalme Intrón | Exón hacen referencia a los dos últimos nucleótidos del intrón y al primer nucleótido del exón aceptor. El porcentaje de los sitios de Exón | Intrón se tomaron independientes a los de Intrón | Exón.

Los alineamientos de cada uno de los ARNm con los genes de las familias de neuropéptidos mostraron que, en LW de hidras las 7 secuencias que presentaron el intrón, este se localizaba al inicio en el extremo 5’. Estas secuencias mostraban nucleótidos adicionales a diferencia de los 20 restantes que no contenían intrones y contaban con una segunda metionina seguida al intrón (Apéndice C).

En las secuencias de AKHs de insectos había un número significativo de secuencias del género *Drosophila,* todas estas presentaron junto con otras pocas secuencias de especies de insectos, un intrón entre el primer y segundo aminoácido del péptido activo (Apéndice D). En 19 secuencias de AKHs de insectos y en 1 de RPCH se evidenció que el gen presentaba una modificación en el codón de la metionina inicial y que se formaba en su mayoría una señal de

AG|G, este cambio provocó que, al alinear el ARNm con el gen, la lectura se originara en otro punto (Apéndice I). Sin embargo, dado que donde se alineó la metionina no existía una señal completa de corte y empalme, no fue posible establecerlo como un límite Exón | Intrón. Debido a la importancia de este resultado, estas secuencias fueron demarcadas dentro de la base de datos con una tonalidad morada para su fácil identificación (Apéndice A, B y K).

El alineamiento de la familia CRZ de invertebrados mostró que 9 secuencias contenían un solo intrón y solamente 1 secuencia presentaba 2 intrones y 3 exones. De las 10 secuencias en total 7 concordaron en la posición del intrón siendo fase 0 y hacia el extremo 3’ y 3 secuencias ubicaron el intrón en el centro del alineamiento. En cuanto a la secuencia que presentaba 2 intrones era de una de las variantes de la especie *Bombyx mori* que presentaba empalme alternativo, la otra variante posee un solo intrón. El intrón que se encuentra ubicado en la posición que no coincide con el resto de secuencias, se podría inferir que es reciente. Sin embargo, se necesitaría de más información experimental y molecular para corroborar y validar esta conjetura. Por otro lado, podría ser solo artefacto del ensamble a nivel genómico o resultado del empalme alternativo.

Adicionalmente, se observó que las 9 secuencias de CRZ de invertebrados se alinearon a partir del segundo exón en relación con la única secuencia que poseía 2 intrones, es decir que para esta variante el primer exón es indicio de una ganancia de ADN a partir del intrón nuevo (Apéndice E).

Con respecto a la GnRH de cordados, esta presentó 76 secuencias y logró poseer la mayor cantidad de intrones con 160 en diferentes fases. Es de resaltar su alto número de intrones dentro de las secuencias a diferencia de otras familias como la AKH de insectos que tenía una mayor cantidad de secuencias 106, pero el número de intrones fue menor con 134 (Tabla 1 y 2, Apéndice H).

La ACP fue la familia que más presentó intrones en fase 2, al considerar el número relativamente bajo de ORFs que se obtuvo en total 29, poseía un poco más del doble de intrones con 62 (Tabla 2). De manera general, la mayoría de sus secuencias presentaron dos intrones y solo algunas hasta 3 (Apéndice B y F). También se encontró indicios de señalización de corte y empalme incompletos dentro de exones en secuencias que estaban distribuidas en todas las familias (Apéndice K).

### 4.3 Correlación de las posiciones de los intrones y límites exónicos de los genes que codifican a las familias: LW, APGW, RPCH, AKH, CRZ, GnRH y sus precursores híbridos con el DNA-LM

Se logró correlacionar con base en los principios y parámetros del DNA-LM, las posiciones de los intrones, los límites exón donador, exón aceptor, fase y número de exones e intrones de las familias LW de hidras, APGW de moluscos, RPCH de crustáceos, AKH de insectos, CRZ de invertebrados, GnRH de cordados y los precursores híbridos ACP y APGW/AKH (Apéndices C-H y K). De esto, se realizó una representación esquemática sintetizada de todas las secuencias que había por familia (Figura 1). La caracterización de la organización de los exones e intrones evidenció que la secuencia del precursor híbrido APGW/AKH (VP-APGW/AKH = HyPro-BpHYR1_030954) contenía 3 exones y 2 intrones, los péptidos activos se ubican en el exón 1 y las 2 copias de la APGW en el exón 2, el primer intrón que se encuentra en sentido 5’ a 3’ es fase 1 y el segundo intrón fase 0. De las 27 secuencias que se obtuvieron de la LW de hidras, solo 7 presentaron intrón, el cual en todas fue fase 2, este estaba ubicado muy cerca al codón de inicio en casi todas las secuencias, de esta manera se observó que las copias de la LW se encontraban en el segundo exón. La única secuencia disponible de la APGW de moluscos y las 3 de RPCH de

crustáceos presentaron un solo intrón en fase 0. Por otro lado, las copias de la APGW como el péptido activo de la RPCH de crustáceos se encontraron en el primer exón. Cabe aclarar que, aunque se haya indicado un solo punto figura 1, en donde, se encuentran las copias de la APGW de moluscos y también de las LW de hidras, estas están distribuidas a lo largo de todo el exón, el punto señalado es una referencia de en qué exón se encuentra.

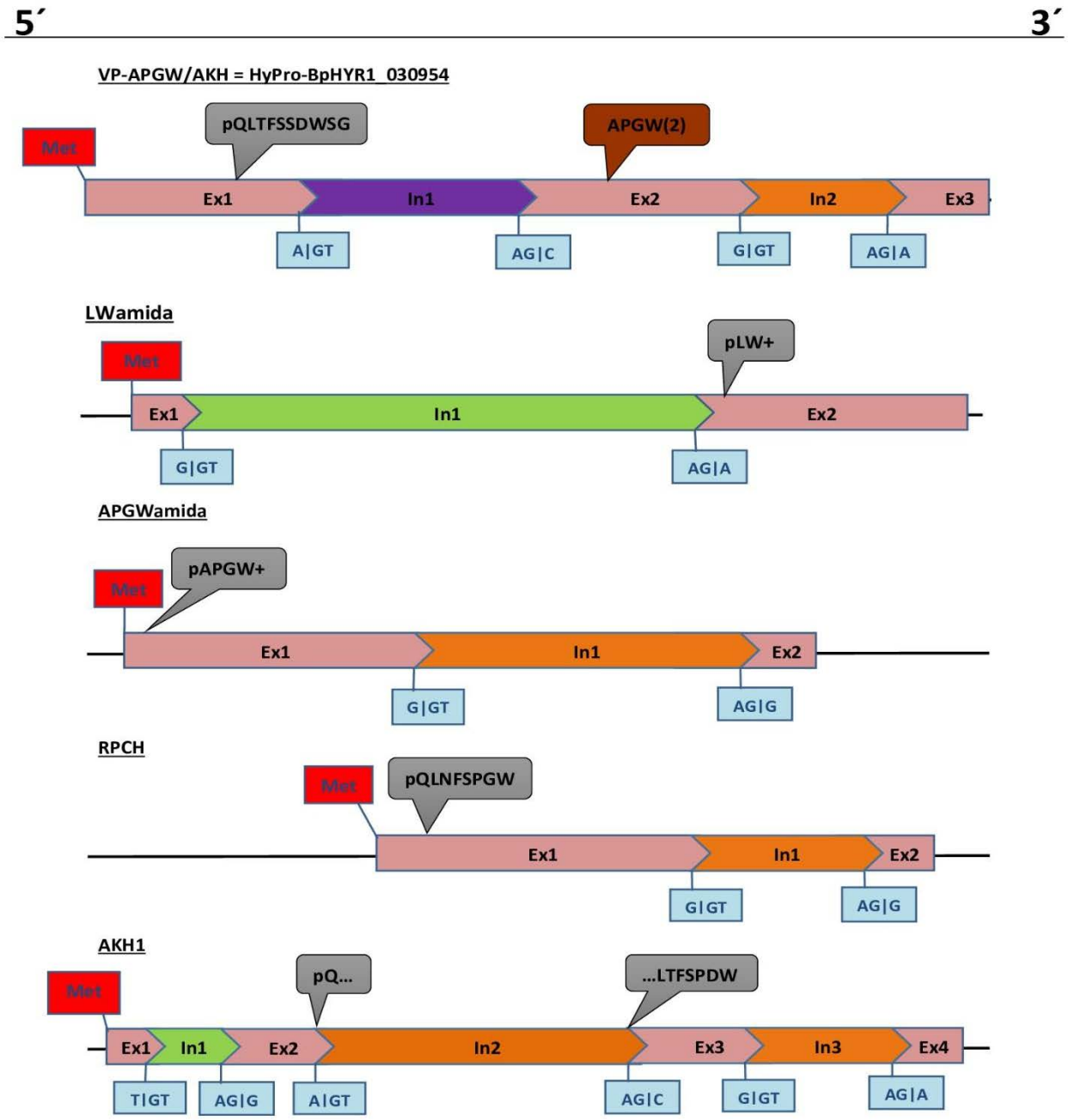
Con respecto a la familia AKH de insectos y sus variantes, se observó que todas comparten la ubicación de un intrón, el cual es fase 0. En la AKH1 y la AKH4 hay un intrón integrado entre el primer y segundo aminoácido del péptido activo; además, la AKH1 y la AKH3 poseen hasta 3 intrones, teniendo 2 ubicados en la misma posición, uno de ellos fase 2 y el otro es el común a todas. La AKH2 es la única que presentó un solo intrón y su ordenamiento es similar al antes descrito para la APGW de moluscos y la RPCH de crustáceos. En el caso la CRZ de invertebrados, se evidenció una estructura parecida a la del precursor híbrido APGW/AKH, aunque el primer intrón que se lee es fase 2 y está en sentido 5'a 3'.

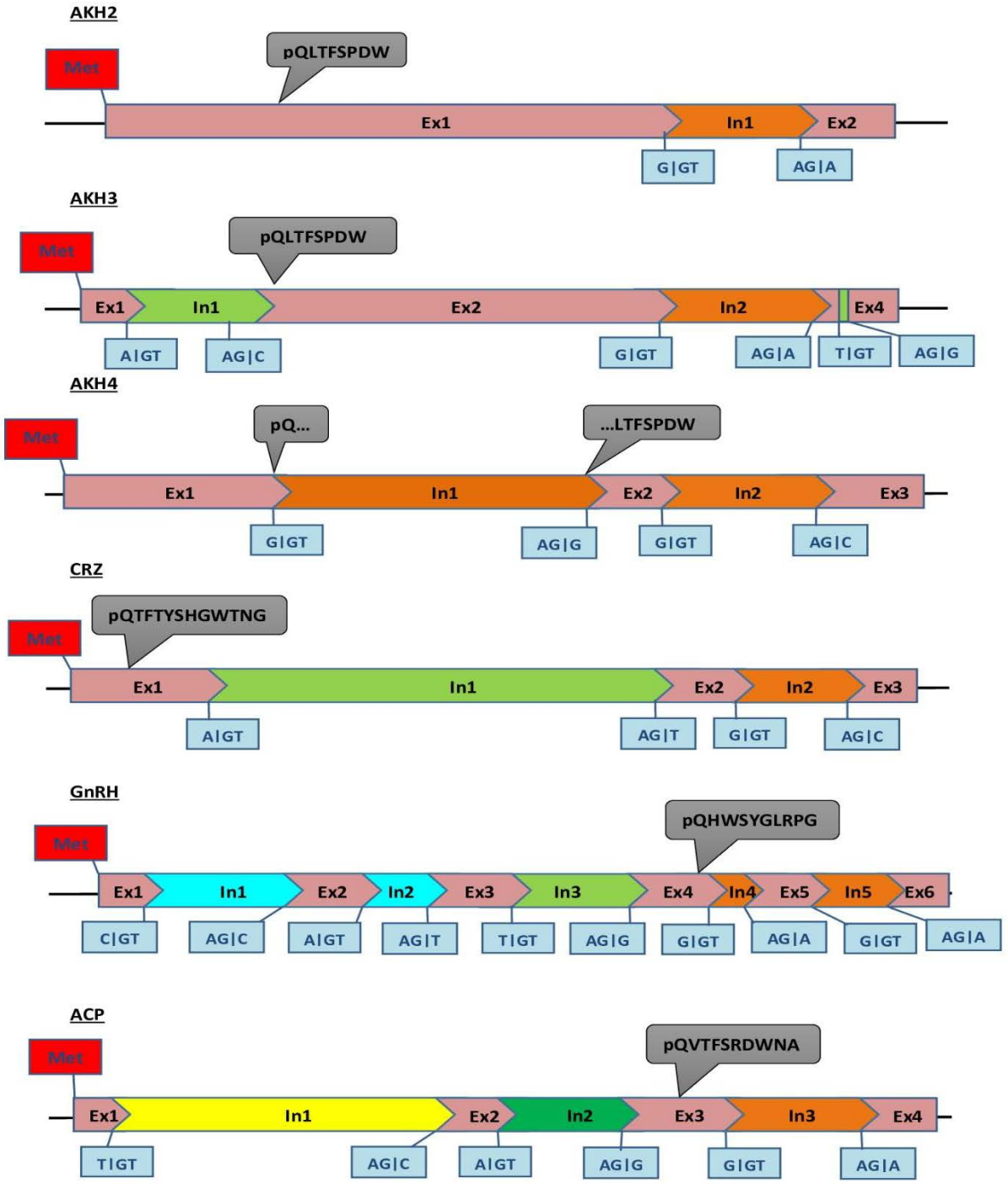
La GnRH de cordados fue la familia con más ganancia de intrones, en su esquema se puede visualizar que 3 de estos son recientes y la conservación de 2 de ellos en más del 50% de las secuencias, estos últimos se localizan muy cerca el uno del otro, en el punto exacto donde se comparte la ubicación del intrón conservado con el resto de las familias; por otra parte, el péptido activo se encontraba en el exón 4. En cuanto a la ACP, fue la única familia con intrones fase 0 a inicio de la secuencia, aunque fue en un porcentaje menor, como también fue la única que mostró que en la mayoría de sus secuencias había un intrón conservado fase 2, su péptido activo se encontraba en el tercer intrón. En general, se obtuvo que todas las familias a excepción de la LW de hidras compartían un intrón conservado en la misma posición, este se ubica hacia el extremo 3', cerca al péptido relacionado, este intrón en todas las secuencias fue fase 0. Con respecto a los

límites Exón | Intrón en su mayoría fueron: G|GT y en los límites Intrón | Exón fue: AG|X. La AKH3 fue la única que presentó un intrón fase 2 en el extremo 3’ y en la ACP se observó, que tenía un intrón conservado fase 2. En conjunto se observa una relación entre la conservación de la posición y la fase del intrón, de esta manera los intrones conservados tienden a ser fase 0 y ubicados hacia el extremo 3’, mientras que los intrones que son recientes tienden a ser fase 1 o 2 y estar localizados cerca al extremo 5’ (Figura 1).

**Figura 1**

*Representación esquemática de la posición de los exones y de los intrones de las familias LW, APGW, RPCH, AKH, CRZ, GnRH y los precursores híbridos ACP y APGW/AKH.*



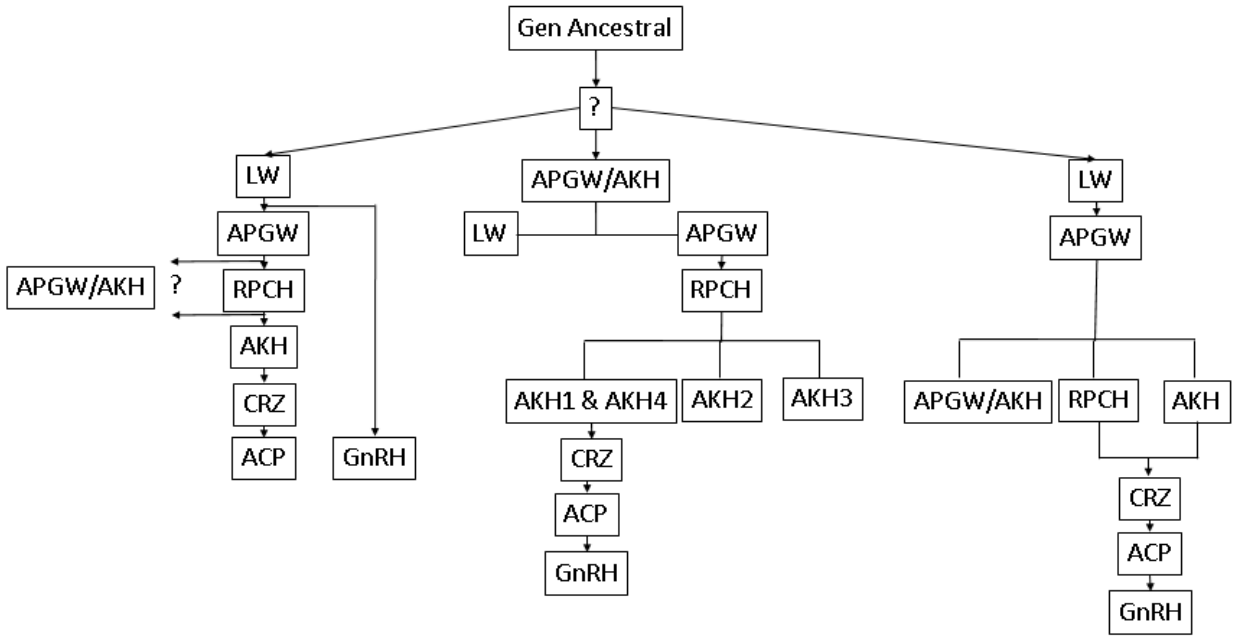

AKH2
Met
pQLTFSPDW
Ex1
In1
Ex2
G|GT
AG|A
AKH3
Met
pQLTFSPDW
Ex1
In1
Ex2
In2
Ex4
A|GT
AG|C
G|GT
AG|A
T|GT
AG|G
AKH4
Met
pQ...
...LTFSPDW
Ex1
In1
Ex2
In2
Ex3
G|GT
AG|G
G|GT
AG|C
CRZ
Met
pQTFTYSHGWTNG
Ex1
In1
Ex2
In2
Ex3
A|GT
AG|T
G|GT
AG|C
GnRH
Met
pQHWSYGLRPG
Ex1
In1
Ex2
In2
Ex3
In3
Ex4
In4
Ex5
In5
Ex6
C|GT
AG|C
A|GT
AG|T
T|GT
AG|G
G|GT
AG|A
G|GT
AG|A
ACP
Met
pQVTFSRDWNA
Ex1
In1
Ex2
In2
Ex3
In3
Ex4
T|GT
AG|C
A|GT
AG|G
G|GT
AG|A
Más del 50% de las secuencias poseen un intron fase 0
Menos del 50% de las secuencias poseen un intron fase 0
Más del 50% de las secuencias poseen un intron fase 1
Menos del 50% de las secuencias poseen un intron fase 1
Más del 50% de las secuencias poseen un intron fase 2
Menos del 50% de las secuencias poseen un intron fase 2

.

*Nota:* La línea negra representa al gen y el cuadro rojo es indicador de en donde empieza la lectura de la secuencia con la Metionina en sentido 5’ a 3’. Los cuadros azules contienen los sitios de proto-empalme, estos fueron posicionados según el porcentaje que presentaron en cada familia. El cuadro gris señala la ubicación del péptido activo, (en el caso de las diferencias en el péptido activo dentro de la familia de las AKHs de insectos, se decidió seleccionar la secuencia del género *Drosophila* a modo de representación general en todas las variedades). Cada color en los intrones atañe a la fase del intrón y el porcentaje de las secuencias que lo presentan en esa posición (la longitud de los exones y de los intrones solo es una representación gráfica sintetizada generada a partir de los alineamientos).

**Figura 2**

*Posibles rutas de evolución de las familias*



*Nota:* Se muestra las posibles formas en las que pudieron evolucionar las familias LW de hidras, APGW de moluscos, RPCH de crustáceos, AKH de insectos, CRZ de invertebrados, GnRH de cordados y los precursores híbridos ACP y APGW/AKH a partir de la relación de sus intrones.

En el lado izquierdo se encuentra la hipótesis inicial de DNA-LM (Martinez-Perez, 2007), en este modelo se propuso un péptido teórico que compartía información entre moluscos y artrópodos. Este péptido podría ubicarse entre la APGW de moluscos y la RPCH de crustáceos o entre la APGW de moluscos y la AKH de insectos. Con el péptido híbrido APGW/AKH se corrobora esta hipótesis al ser encontrado en la naturaleza en la especie *B. plicatilis*. En el modelo del centro se muestra una evolución a partir del precursor híbrido APGW/AKH, indicando una evolución a partir de los primeros eucariontes, en este se observa que la LW de hidras se formó independientemente de la APGW de moluscos. La LW DE HIDRAS se desarrolla a partir del exón 2 con la pérdida de los intrones ancestrales, mientras que la APGW de moluscos pierde uno y conserva otro en la posición 3’ que hereda al resto de familias; además de que se visualiza la variación de las AKHs de insectos a partir del movimiento de intrones, en esta hipótesis la CRZ de invertebrados, la ACP provienen de la AKH1 y la AKH4. El tercer modelo muestra una evolución más ramificada, el precursor híbrido APGW/AKH se genera a partir de la APGW de moluscos, junto a las RPCH de crustáceos y AKH de insectos, estas últimas aumentaron el ingreso de intrones, dando al resto de familias.

## 5. Discusiones

En esta pasantía se contribuyó al conocimiento del DNA-LM mediante el análisis de intrones en genes que codifican para los precursores neuropéptidicos, lo cual, es uno de los dos elementos que sustentan el modelo evolutivo. En un principio la investigación de Cadena-Caballero (2020), realizó un estudio filogenético entre las familias de neuropéptidos LW de hidras, APGW de moluscos, RPCH de crustáceos, AKH de insectos, CRZ de invertebrados, GnRH de cordados y sus precursores híbridos empleando una base de datos que contenía 878 secuencias aminoacídicas, un gran porcentaje de estas fueron obtenidas por medio de su ADN complementario (ADNc) a partir de proyectos de transcriptomas en el LGCA. Sin embargo, dado que la investigación tenía un enfoque a nivel aminoacídico, no se tuvo en cuenta el gen correspondiente para cada secuencia. Por lo tanto, al realizar la depuración con el paramento de secuencias que presentaban el gen en la base de datos del GenBank y del descarte de versiones obsoletas, únicamente se obtuvo un 28.8% de la base original correspondiente a los genes y genomas reportados. Aun con la falta de secuencias de genes de todos los neuropéptidos reportados a nivel de ADNc, en esta pasantía se logró contribuir al DNA-LM desde la perspectiva de los ácidos nucleicos y aportar en especial relevancia al movimiento de intrones para su evolución. No obstante, a consecuencia de la desigualdad en el número de genes reportados que se presentó entre familias, lo cual, es generado a que una gran cantidad de las especies empleadas en proyectos de investigación son modelos de estudio biológico dado el interés comercial o de otras índoles, conlleva que para este trabajo los resultados fueran sesgados hacia familias con mayor número de secuencias de estas especies como, por ejemplo, la AKH de insectos y la GnRH de cordados.

A pesar de esta limitante se estableció que los dinucleótidos presentes en los bordes del intrón, en su mayoría corresponden a las secuencias canónicas |GT…AG|, de acuerdo con lo previamente mencionado por otros autores (Parada et al., 2014; Pucker y Brockington 2018), lo que sugiere que no solo se está heredando la posición de los intrones, sino también los límites intrónicos. De lo anterior, podría inferirse una posible correlación entre la conservación de la posición y la de los límites intrónicos (Pucker y Brockington 2018).

En cuanto al porcentaje minoritario de los límites intrónicos |GC…AG| que hace referencia a 4 intrones que estaban dentro de secuencias de GnRH de cordados (1 de *Oncorhynchus mykiss* y 3 dentro de 3 secuencias de la especie *Salmo salar*), y también de 2 intrones que estaban en ACP (en 2 secuencias de la especie *Bicyclus anynana)*, esta secuencia no canónica se encontró que también es reconocida como una señal de corte de intrones por la subunidad U2 del espliceosoma, solo que es menos frecuente encontrarla (Poverennaya y Roytberg, 2020). El porcentaje en los límites |GT…TG| en 2 secuencias de ACP de lepidóptero *Helicoverpa armigera*, fueron grandes hallazgos dado que esta secuencia no canónica ha sido reportada en algunos genomas de mamíferos, y aunque los autores obtuvieron igualmente un bajo porcentaje, concluyen que debido a las imperfecciones en las anotaciones realizadas en genomas tan amplios, el porcentaje minoritario en mamíferos estaba siendo subestimado y que había una gran probabilidad de que fuera mayor (Alioto, 2007; Poverennaya y Roytberg, 2020), lo cual, lleva a cuestionar si su presencia en grupos aislados se deba a algún grado de conservación o si este resultado es de un proceso aleatorio.

Así, se analizó en qué posición se encontraban estos dos intrones y se encontró que coincidían con la ubicación del intrón ortólogo de la mayoría de secuencias en todas las familias. Sin embargo, esta idea se contrarresta a la posible correlación entre la conservación del sitio de

proto-empalme y la de la posición del intrón, ya que se esperaría que intrones conservados en su ubicación mantuvieran a su vez la secuencia canónica |GT...AG| (Parada et al., 2014; Pucker y Brockington 2018). De esta forma, se consideró que era más probable que su bajo porcentaje se tratase de una mutación de la purina A por la pirimidina T, sin embargo, esta secuencia sigue siendo reconocida por el espliceosoma como un corte intrónico, lo que sugiere que puede estar mediada por mecanismos y/o variantes genéticas que generan sitios no canónicos (Riepe et al., 2021).

En cuanto a los límites exónicos, se determinó que en general los nucleótidos variaban en sus bordes. Lo que entra en contraste con lo que sugieren Long et al (1998), donde expresan que existe un grado de conservación en los nucleótidos del exón que flaquean con el intrón, y que es resultado de la remanencia de sitios de proto-empalme en eucariotas ancestrales que sirvieron para el ingreso de nuevos intrones. De ser así, debería existir una secuencia canónica establecida para los límites exónicos, en cambio, lo que se observa en los estudios de análisis de intrones es que existe una tendencia en los sitios de proto-empalme a mantener conservado los límites intrónicos en vez de los límites exónicos (Lynch., 2002; Roesner et al., 2005; Csuros et al., 2011; Parenteau et al., 2011; Sêton et al., 2016; Mukhopadhyay y Hausner., 2021), a ejemplo de esto, resulta la secuencia canónica de intrones |GT…AG|, ya establecida y que se encuentra en el 98% de las secuencias conservada (Parada et al., 2014; Pucker y Brockington 2018).

La conformación de los intrones y exones dentro de las familias fue indicio del movimiento de intrones en su evolución. La secuencia del precursor híbrido APGW/AKH (VP- APGW/AKH = HyPro-BpHYR1_030954), pertenece a la especie *Brachionus plicatilis*, un rotífero cuya organización intrón/exón coincide con la de uno de los genes ancestrales propuestos hace 20 años por Martínez-Pérez et al (2002), quienes postularon un gen con una copia de algún miembro de la

AKH de insectos o RPCH de crustáceos con copias de APGW de moluscos. Es decir, un gen neuropéptidico híbrido de moluscos y artrópodos. En esta pasantía, se demuestra que el gen si está presente en la naturaleza y está conformado por 2 intrones y 3 exones, la posición del intrón que estaba hacia el extremo 3’ coincide en la ubicación del intrón conservado del resto de familias, lo que afirma más la relación ancestral que hay entre las familias evaluadas.

La presencia de la APGW/AKH en el genoma de *B. plicatilis* reconstruye una representación innovadora a los procesos evolutivos de genes neuroendocrinos sobre la ocurrencia en la transición de genes de especies eucariotas simples a otras con mayor complejidad (Martínez-Pérez et al., 2007). En este sentido, es de resaltar que debido al bajo porcentaje de secuencias que posee el intrón fase 2 al inicio de la secuencia en LW de hidras, la cual, se propuso en el DNA-LM que perdió el codón de Leucina para generar a la APGW de moluscos (Martínez-Pérez et al., 2007); en función de los resultados obtenidos ahora se podría proponer que este hecho es una evidencia de ser un intrón reciente y dado que las secuencias que lo presentan mostraron un aumento de nucleótidos, se deduce que se genera ganancia de ADN en el extremo 5’.

En cuanto a la ausencia del intrón ortólogo en las secuencias de LW de hidras, presente en el resto de las familias y a la falta de factores de empalme dentro del ORF, se podría proponer que la evolución independiente de esta familia se formó a partir del segundo exón, el cual, contiene las copias de APGW, en este sentido se habría perdido dominios del primer y tercer exón. A partir de esto se podría postular que, aun cuando la LW de hidras comparte homología con estas familias, la formación de la APGW de moluscos no necesariamente provino de la LW de hidras, como se afirma en uno de los principios del DNA-LM (Martínez-Pérez et al 2007), sino que evolucionaron separadamente. Sin embargo, el hecho de que el precursor AKH/APGW tenga los péptidos activos

en cada uno de sus exones abre la posibilidad de proponer diversas hipótesis, las cuales deberán ser validadas con los precursores de neuropéptidos de otros eucariontes y de manera experimental.

En el caso de la APGW de moluscos esta perdió dominios del primer exón a causa de la deleción del primer intrón, los fragmentos del primer exón se unieron al segundo, lo que generó un aumento de las copias de APGW y una conservación del tercer exón (Martínez-Pérez et al., 2002). Posteriormente, se formó la RPCH de crustáceos a través de pérdidas del ADN en lo que sería ahora su primer exón, estructura que se conservó hasta cierto grado en las AKHs de insectos, se cree que la variedad que se ha dado en esta familia está dada por el movimiento de intrones, esto coincide con lo propuesto inicialmente en el DNA-LM que indica que la diferencia de posiciones de los intrones dentro del gen pudo dar estas diversificaciones dentro de la familia AKH de insectos (Martínez-Pérez et al., 2002; Martínez-Pérez et al., 2007). La presencia de un segundo intrón fase 0 integrado en medio del péptido activo en un gran porcentaje de AKH1 y AKH4 puede deberse a una mejora en el proceso de recombinación homóloga (Fedorova et al., 2003). O dado que la mayoría de las especies de la AKH1 que lo presentaban eran del género *Drosophila*, las cuales, no presentaban el intrón ortólogo en la ubicación que concuerda con el resto de las familias, se postula que podría darse un “deslizamiento de intrones”, lo que produjo que el intrón conservado se situará aguas arriba hacia el extremo 5’ y que esto fuera replicado en especies cercanas filogenéticamente o heredado de la AKH1 a la AKH4 (Lehmann et al., 2010).

El porcentaje de ganancia de intrones presente en GnRH de cordados, podría estar relacionado con la longitud de sus genes, algunos autores explican que genes que contienen abundancia en sus pares de bases, tienen una relación directamente proporcional con el número de intrones por secuencia (Hadrill et al., 2005; Keane y Seoighe, 2016). La presencia del intrón conservado en la GnRH de cordados y que comparte con el resto de familias favorece la hipótesis

del DNA-LM, ya esta se separó evolutivamente antes de la formación de la APGW de moluscos (Cadena-Caballero, 2020), lo que resulta interesante teniendo en cuenta que comparten este intrón, lo que sugiere que si existe una relación ancestral entre estas familias al comparar las relaciones filogenéticas con nuestros resultados (Hauser y Grimmelikhuijzen, 2014; Plachetzki et al., 2016). Un punto a destacar a consecuencia del ingreso de intrones, es el movimiento del péptido activo, si bien en la mayoría de familias este se ubica en el exón 1, al observarse familias como GnRH de cordados y ACP que presentan más intrones a lo largo de la secuencia y teniendo en cuenta que los intrones recientes se observaron en su mayoría hacia el extremo 5', el péptido activo tendría una tendencia a irse hacia el extremo 3' para estas dos familias, estudios experimentales posteriores son necesarios para validar si hay alguna repercusión a nivel funcional o evolutivo sobre este.

El intrón conservado fase 2 que posee las ACP, se encuentra ubicado en la misma posición que el intrón conservado fase 0 de la AKH1, la AKH4 y el intrón reciente fase 2 de la CRZ de invertebrados, el péptido activo en ACP se encuentra similar al del péptido activo de la GnRH de cordados, lo que afirmaría la relación filogenética entre estas familias, dado que la ACP es un precursor híbrido que presenta características provenientes de la AKH de insectos, CRZ de invertebrados y GnRH de cordados (Hansen et al., 2010).

La presencia de sitios remanentes de proto-empalme distribuidos dentro del ORF en algunas secuencias, permite inferir una posible pérdida de intrones en esas ubicaciones, Martínez-Pérez et al (2002) explica que la disposición de potenciadores de empalme en secuencias de genes que carecen de intrones es indicio de un posible intrón presente en un gen ancestral que con él tiempo se perdió. Sin embargo, no se observó un patrón en la ubicación de estos, ni tampoco fue significativa el número de secuencias que los poseían, por lo que se descarta la idea de que fueran intrones ortólogos. A partir de esto se deduce para este estudio que se debe a una posible falla en

el proceso de integración de nuevos intrones que no logran establecerse del todo y que se terminan perdiendo, esto podría darse por una mutación en el codón en el cual está insertado el intrón (Riepe et al., 2021).

La recombinación homóloga también podría ser responsable, ya que durante el apareamiento de los cromosomas en la fase de zigoteno, el cruce podría interferir en las señales que definen al intrón, por consiguiente, el espliceosoma no lograría identificarlo y escindirlo. Así generaría una ganancia de fragmentos de ADN, una señalización de corte y empalme defectuosa (Fedorova et al., 2003). No obstante, esta idea es poco probable ya que la recombinación homóloga ha sido asociada con el *splicing* para generar un proceso más eficiente (Fedorova et al., 2003).

Dado los resultados obtenidos, se logró relacionar la fase y posición del intrón con el DNA-LM en secuencias que presentaron bajos porcentajes de intrones al inicio de la lectura y también de las secuencias demarcadas en tono morado de las AKHs de insectos y RPCH de crustáceos (Figura 1, Apéndice B y I), en este sentido se observó una cantidad muy baja de secuencias que presentaban algunos intrones hacia el extremo 5’ y que estos estaban parcializados a ser fase 1 o 2. Mientras que el intrón ortólogo presente en la mayoría de secuencias en todas las familias se encontraba hacia el extremo 3’ siendo en todos los casos fase 0. Baulin et al (2020), explica que la presencia de intrones fase 1 al inicio de la secuencia no es aleatoria, especialmente en genes relacionados con el cerebro y que esta fase está relacionada con la péptida señal y la longitud del intrón. Sin embargo, ese estudio no tuvo en cuenta los intrones de fase 2 y 0 ni su posición y si coinciden con la de intrones fase 1.

Por otra parte, la inclinación por un porcentaje mayor de intrones fase 0 ha sido explicada de diferentes aspectos, uno de ellos expone una relación con el sitio de proto-empalme más frecuente G|G, alegando que es más probable encontrar esta secuencia entre codones que dentro

de los mismos (Ruvinsky et al., 2005). La hipótesis de intrones tempranos explica tal cuestión, argumentando que los intrones ancestrales eran fase 0 y que estos provienen de fracciones de genes de procariotas que los heredaron a los eucariotas y que a lo largo del tiempo solo se ha aumentado su proporción (de Sousa et al., 1998; Roy, 2003). Más aún, ninguna de estas propuestas explica la relación de las fases con su patrón de disposición. Sin embargo, aquellos autores que apoyan la hipótesis de intrones tardíos, expresan que la falta de uniformidad de la distribución de las fases se debe a que no es aleatoria la integración de nuevos intrones (Nguyen et al., 2006). Nuestros resultados concuerdan con esta última premisa siendo clave.

En relación con los intrones recientes de las familias se evidenció cierto grado de "inestabilidad" en el codón de la metionina, ya que este tiende a cambiar su segundo nucleótido de una Timina por una Guanina. De esta forma, se desarrolla una señal de corte y empalme AG|G y se observa que la mayoría de los intrones recientes fueron fase 1 y/o 2, y que el cambio se ve dentro del codón lo que produce la señalización, y así mismo una inserción entre el primer y segundo nucleótido (fase 1) o segundo y tercer nucleótido (fase 2).

No obstante, se cree que los intrones fase 1 o 2 no logran establecerse del todo, esto considerando que los intrones que son conservados en grupos filogenéticamente lejanos y cercanos son de fase 0, lo que se cree que puede haber una tendencia a volverse a esta fase. Esto es apoyado por Nguyen et al (2006), donde afirma que no existe una uniformidad en la proporción de las fases, con una frecuencia mayor de intrones fase 0.

Continuando con la idea de la propensión de los intrones a volverse fase 0, se cree que en intrones fase 1 y 2 se generaría un "movimiento" en la secuencia, en otras palabras, cuando un intrón logra establecerse entre dos codones, se vuelve más estable en su posición, y por consiguiente es más probable a que se conserve (Nguyen, 2006). Este movimiento puede estar

relacionado con el llamado “deslizamiento de intrones” (*Intron sliding*), básicamente es el desplazamiento de los límites exónicos e intrónicos en una distancia corta, esto ocurriría por medio del empalme alternativo, originalmente esta idea surgió por parte de los partidarios de la hipótesis de intrones tempranos (Tarrio et al., 2008), a modo de explicar porque variaron las posiciones de los intrones ortólogos, posteriormente, los que apoyan la hipótesis de intrones tardíos, no negaron la posibilidad de esta idea, sin embargo declararon que si ocurre, no tendría un efecto relevante en la diversidad de la posición (Tarrio et al., 2008). Para nuestros resultados creemos que las posiciones de los intrones ortólogos pueden variar, pero en una distancia poco significativa. En cuanto a las fases, se alude que el deslizamiento de intrones podría influenciar el cambio de una a otra; lo que se observa en las secuencias de *Homo sapiens* y de *Macaca mulatta* de GnRH que presentan un deslizamiento del intrón entre 3 a 6 aminoácidos hacia el extremo 3’ siendo fase 1, en comparación con la ubicación del intrón conservado fase 0 del resto de secuencias.

Este argumento está respaldado por el trabajo de Poverennaya et al (2017), donde se describe una posible conexión entre el cambio de fase y el deslizamiento de intrones. Esto es corroborado también en un estudio comparativo de la ubicación de intrones en genes de familias diferentes y se asegura que es muy posible el evento evolutivo en que el intrón se desplace de a un solo nucleótido (Rogozin et al., 2000).

A partir de lo anterior, se puede inferir que la modificación en la metionina ocasiona que el inicio de la lectura ya no se produzca a partir de ese punto, en tal caso puede ocurrir una de dos situaciones: en la primera, que exista otra metionina río abajo, en este punto, se cree que habría una pérdida de fragmentos de ADN. En tanto, si sucede la segunda situación donde la lectura empiece con una metionina aguas más arriba, bajo este caso, esta metionina debería presentar una señalización Exón|Intrón para que sea reconocido por el espliceosoma y así se genere un exón

demás, con esta condición, habría ganancia de nucleótidos. En cualquiera de los dos casos, dicha modificación afectaría la secuencia, cuyos efectos formarían nuevos dominios, siendo indicio de la evolución de estas familias de neuropéptidos como lo postula DNA-LM.

Finalmente, estos resultados son cruciales para redefinir los principios, parámetros y alcances del DNA-LM ya que se observó que la evolución de estas familias neuropeptídicas no solo estuvo dada por la pérdida de ADN, sino que el ingreso de nuevos intrones genera también una ganancia de ácidos nucleicos como se observó en algunas familias. Es decir, el resultado del movimiento de intrones en las familias neuropeptídicas permite el desarrollo de nuevos dominios moleculares que pueden darse tanto por pérdida como por ganancia de intrones. Por otra parte, la presencia de intrones dentro de la secuencia APGW/AKH del rotífero *B. plicatilis* abre la posibilidad de estudiar el modelo desde especies con una organización celular más sencilla y no solo en nuevas especies de invertebrados como se había descrito inicialmente en el DNA-LM (Martínez-Pérez et al., 2002; Martínez-Pérez et al., 2007; Cadena-Caballero, 2020).

## 6. Conclusiones

La organización y caracterización de los intrones y exones en las secuencias permite tener un indicio de cómo se está dando la evolución de las familias LW de hidras, APGW de moluscos, RPCH de crustáceos, AKH de insectos, CRZ de invertebrados, GnRH de cordados y precursores híbridos, a través del movimiento de intrones, tal como lo postula el DNA-LM.

La visible conservación de los límites intrónicos y la variedad en los límites exónicos, pone en relevancia la prevalencia de las secuencias canónicas de los intrones en el proceso de reconocimiento del espliceosoma, efectuando una importancia mayor en el mecanismo de definición del intrón que en el de definición del exón.

El patrón de distribución de las fases, posicionando al intrón conservado, siendo fase 0 en el extremo 3’ y a los intrones nuevos siendo fase 1 y 2, ubicados en el extremo 5’, permite apoyar la idea propuesta de origen de intrones tardíos, afirmando que esta correlación se debe a que los sitios de proto-empalme no se forman aleatoriamente, y por ende tampoco el ingreso de nuevos intrones.

La conservación de un intrón en la mayoría de las familias logra establecerlo como “ortólogo”, lo que deja ver la relación ancestral que existe entre estas, en el caso de la LW de hidras pudo haberlo perdido en su formación.

La presencia de intrones en proporciones bajas en las secuencias, integrados en su mayoría en el extremo 5’, y del notable desplazamiento del péptido activo por familia, fue evidencia de la evolución independiente en cada linaje.

## 7. Recomendaciones

Se recomienda mayores estudios en torno a la evolución de neuropéptidos con mayor número de genes que permita generar resultados más certeros. Así mismo ampliar el apoyo a sistemas de secuenciación de nueva generación para obtener transcriptomas de especies no solo modelo sino de la vida silvestre o de interés comercial. Se propone aplicar el DNA-LM a organismos aún inexplorados como parte del proceso de evaluación y caracterización de las diversas hipótesis. Adicionalmente, de la creación y el desarrollo de software bioinformáticos aplicado a la supercomputación, el cual, permite analizar de manera fácil, rápida y certera conjuntos de datos masivo de genes que codifican para neuropéptidos con la identificación de sus intrones y exones.

## Referencias Bibliográficas

Alioto, T. S. (2007). U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Research, 35*(Database), D110–D115. https://doi.org/10.1093/nar/gkl796

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Baulin, E. F., Kulakovskiy, I. V., Roytberg, M. A., & Astakhova, T. V. (2020). Brain-related genes are specifically enriched with long phase 1 introns. *PloS one*, *15*(5), e0233978. https://doi.org/10.1371/journal.pone.0233978

Cadena-Caballero, C. E. (2020). *Contribution to the DNA loss model for the study of the evolutionary relationship of neuropeptide families: LWamide, APGWamide, Red Pigment Concentrating Hormone, Adipokinetic Hormone, Corazonin and Gonadotropin-Releasing Hormone*. Bachelor thesis. [*Industrial University of Santander*]. https://doi.org/10.13140/RG.2.2.35621.63208

Chorev, M., & Carmel, L. (2012). The Function of Introns. *Frontiers in Genetics*, *3*(4), 1–15. https://doi.org/10.3389/fgene.2012.00055

Chow, L. T., Gelinas, R. E., Broker, T. R., & Roberts, R. J. (1977). An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell, 12*(1), 1–8. https://doi.org/10.1016/0092-8674(77)90180-5

Csuros, M., Rogozin, I. B., & Koonin, E. V. (2011). A Detailed History of Intron-rich Eukaryotic Ancestors Inferred from a Global Survey of 100 Complete Genomes. *PLoS Computational*

*Biology*, *7*(9), e1002150. https://doi.org/10.1371/journal.pcbi.1002150

Creer S. (2007). Choosing and using introns in molecular phylogenetics. *Evolutionary bioinformatics online*, (*3*), 99–108.

de Souza, S. J., Long, M., Klein, R. J., Roy, S., Lin, S., & Gilbert, W. (1998). Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins. *Proceedings of the National Academy of Sciences*, *95*(9), 5094–5099. https://doi.org/10.1073/pnas.95.9.5094

Dibb, N. J., & Newman, A. J. (1989). Evidence that introns arose at proto-splice sites. *The EMBO Journal*, *8*(7), 2015–2021. https://doi.org/10.1002/j.1460-2075.1989.tb03609.x

Fedorova, L., & Fedorov, A. (2003). Introns in gene evolution. *Genetica*, *118*(1), 123–131. https://doi.org/10.1023/A:1024145407467

Forsdyke, D. R. (2013). Introns First. *Biological Theory*, *7*(3), 196–203. https://doi.org/10.1007/s13752-013-0090-6

García-López, M. J., Martínez Martos, J. M., Mayas Torres, M. D., Carrera González, M. P., & Ramírez Expósito, M. J. (2002). Physiology of the neuropeptides. *Revista de Neurología*, *35*(08), 784. https://doi.org/10.33588/rn.3508.2002028

Gilbert, W. (1978). Why genes in pieces? *Nature*, *271*(5645), 501–501. https://doi.org/10.1038/271501a0

Gilbert, W. (1987). The Exon Theory of Genes. *Cold Spring Harbor Symposia on Quantitative Biology*, *52*, 901–905. https://doi.org/10.1101/SQB.1987.052.01.098

Gilbert, W., & Glynias, M. (1993). On the ancient nature of introns. *Gene*, *135*(1–2), 137–144. https://doi.org/10.1016/0378-1119(93)90058-B

Giovannoni, S. J., Cameron Thrash, J., & Temperton, B. (2014). Implications of streamlining

theory for microbial ecology. *The ISME Journal*, *8*(8), 1553–1565. https://doi.org/10.1038/ismej.2014.60

Haddrill, P. R., Charlesworth, B., Halligan, D. L., & Andolfatto, P. (2005). Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content. *Genome Biology*, *6*(8), 1–8. https://doi.org/10.1186/gb-2005-6-8-r67

Hansen, K. K., Stafflinger, E., Schneider, M., Hauser, F., Cazzamali, G., Williamson, M., Kollmann, M., Schachtner, J., & Grimmelikhuijzen, C. J. P. (2010). Discovery of a Novel Insect Neuropeptide Signaling System Closely Related to the Insect Adipokinetic Hormone and Corazonin Hormonal Systems. *Journal of Biological Chemistry*, *285*(14), 10736–10747. https://doi.org/10.1074/jbc.M109.045369

Hauser, F., & Grimmelikhuijzen, C. J. (2014). Evolution of the AKH/corazonin/ACP/GnRH receptor superfamily and their ligands in the Protostomia. *General and comparative endocrinology*, 209, 35–49. https://doi.org/10.1016/j.ygcen.2014.07.009

Jiménez-García, E., Tapia-Vieyra, J. V., & Mas-Oliva, J. (2004). El esplaiceosoma: Corte y emplame del Pre-ARNm. *REB*, *23*(2), 59–63.

Jo, B.-S., & Choi, S. S. (2015). Introns: The Functional Benefits of Introns in Genomes. *Genomics & Informatics*, *13*(4), 112. https://doi.org/10.5808/GI.2015.13.4.112

Keane, P. A., & Seoighe, C. (2016). Intron Length Coevolution across Mammalian Genomes. *Molecular Biology and Evolution*, *33*(10), 2682–2691. https://doi.org/10.1093/molbev/msw151

Koonin, E. V. (2006). The origin of introns and their role in eukaryogenesis: A compromise solution to the introns-early versus introns-late debate? *Biology Direct*, *1*(1), 1–23. https://doi.org/10.1186/1745-6150-1-22

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*(21), 2947–2948. https://doi.org/10.1093/bioinformatics/btm404

Lehmann, J., Eisenhardt, C., Stadler, P. F., & Krauss, V. (2010). Some novel intron positions in conserved Drosophila genes are caused by intron sliding or tandem duplication. *BMC Evolutionary Biology*, *10*(1), 156. https://doi.org/10.1186/1471-2148-10-156

Lim, C. S., T. Wardell, S. J., Kleffmann, T., & Brown, C. M. (2018). The exon–intron gene structure upstream of the initiation codon predicts translation efficiency. *Nucleic Acids Research*, *46*(9), 4575–4591. https://doi.org/10.1093/nar/gky282

Logsdon, J. M. (1998). The recent origins of spliceosomal introns revisited. *Current Opinion in Genetics & Development*, *8*(6), 637–648. https://doi.org/10.1016/S0959-437X(98)80031-2

Long, M., De Souza, S. J., Rosenberg, C., & Gilbert, W. (1998). Relationship between "proto-splice sites" and intron phases: Evidence from dicodon analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(1), 219–223. https://doi.org/10.1073/pnas.95.1.219

Lynch, M. (2002). Intron evolution as a population-genetic process. *Proceedings of the National Academy of Sciences*, *99*(9), 6118–6123. https://doi.org/10.1073/pnas.092595699

Martínez-Pérez, F., Becerra, A., Valdés, J., Zinker, S., & Aréchiga, H. (2002). A possible molecular ancestor for mollusk APGWamide, insect Adipokinetic Hormone, and crustacean Red Pigment Concentrating Hormone. *Journal of Molecular Evolution*, *54*(6), 703–714. https://doi.org/10.1007/s00239-001-0036-7

Martínez-Pérez, F., Durán-Gutiérrez, D., Delaye, L., Becerra, A., Aguilar, G., & Zinker, S. (2007).

Loss of DNA: A plausible molecular level explanation for crustacean neuropeptide gene evolution. *Peptides*, *28*(1), 76–82. https://doi.org/10.1016/j.peptides.2006.09.021

Martínez-Pérez, F., Bendena, W. G., Chang, B. S., & Tobe, S. S. (2009). FGLamide Allatostatin genes in Arthropoda: introns early or late?. *Peptides*, *30*(7), 1241–1248. https://doi.org/10.1016/j.peptides.2009.04.001

Mukhopadhyay, J., & Hausner, G. (2021). Organellar Introns in Fungi, Algae, and Plants. *Cells*, *10*(8), 2001. https://doi.org/10.3390/cells10082001

Nguyen, H. D., Yoshihama, M., & Kenmochi, N. (2006). Phase distribution of spliceosomal introns: implications for intron origin. *BMC Evolutionary Biology*, *6*(1), 69. https://doi.org/10.1186/1471-2148-6-69

Parada, G. E., Munita, R., Cerda, C. A., & Gysling, K. (2014). A comprehensive survey of non-canonical splice sites in the human transcriptome. *Nucleic Acids Research*, *42*(16), 10564–10578. https://doi.org/10.1093/nar/gku744

Parenteau, J., Durand, M., Morin, G., Gagnon, J., Lucier, J.-F., Wellinger, R. J., Chabot, B., & Abou Elela, S. (2011). Introns within Ribosomal Protein Genes Regulate the Production and Function of Yeast Ribosomes. *Cell*, *147*(2), 320–331. https://doi.org/10.1016/j.cell.2011.08.044

Plachetzki, D. C., Tsai, P. S., Kavanaugh, S. I., & Sower, S. A. (2016). Ancient origins of metazoan gonadotropin-releasing hormone and their receptors revealed by phylogenomic analyses. *General and comparative endocrinology*, *234*, 10–19. https://doi.org/10.1016/j.ygcen.2016.06.007

Poverennaya, I. V., Gorev, D. D., Astakhova, T. V., Tsitovich, I. I., Yakovlev, V. V., & Roytberg, M. A. (2017). Intron Sliding and Length Variability of Genes Enriched of Phase 1 Long

Introns. *Mathematical Biology and Bioinformatics*, *12*(2), 302–316. https://doi.org/10.17537/2017.12.302

Poverennaya, I. V., & Roytberg, M. A. (2020). Spliceosomal Introns: Features, Functions, and Evolution. *Biochemistry (Moscow)*, *85*(7), 725–734. https://doi.org/10.1134/S0006297920070019

Pucker, B., & Brockington, S. F. (2018). Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. *BMC Genomics*, *19*(1), 980. https://doi.org/10.1186/s12864-018-5360-z

Riepe, T. V., Khan, M., Roosing, S., Cremers, F., & 't Hoen, P. (2021). Benchmarking deep learning splice prediction tools using functional splice assays. *Human mutation*, 42(7), 799–810. https://doi.org/10.1002/humu.24212

Roesner, A., Fuchs, C., Hankeln, T., & Burmester, T. (2005). A Globin Gene of Ancient Evolutionary Origin in Lower Vertebrates: Evidence for Two Distinct Globin Families in Animals. *Molecular Biology and Evolution*, *22*(1), 12–20. https://doi.org/10.1093/molbev/msh258

Rogozin, I. B., Carmel, L., Csuros, M., & Koonin, E. V. (2012). Origin and evolution of spliceosomal introns. *Biology Direct*, *7*(1), 11. https://doi.org/10.1186/1745-6150-7-11

Rogozin, I. B., Lyons-Weiler, J., & Koonin, E. V. (2000). Intron sliding in conserved gene families. *Trends in Genetics*, *16*(10), 430–432. https://doi.org/10.1016/S0168-9525(00)02096-5

Roy, S. W. (2003). Recent evidence for the Exon Theory of Genes. *Genetica*, *118*(2–3), 251–266. https://doi.org/10.1023/A:1024190617462

Ruvinsky, A., Eskesen, S. T., Eskesen, F. N., & Hurst, L. D. (2005). Can Codon Usage Bias

Explain Intron Phase Distributions and Exon Symmetry? *Journal of Molecular Evolution*, *60*(1), 99–104. https://doi.org/10.1007/s00239-004-0032-9

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., & Karsch-Mizrachi, I. (2019). GenBank. *Nucleic Acids Research*, *48*(1), 84–86. https://doi.org/10.1093/nar/gkz956

Sêton Bocco, S., & Csűrös, M. (2016). Splice Sites Seldom Slide: Intron Evolution in Oomycetes. *Genome Biology and Evolution*, *8*(8), 2340–2350. https://doi.org/10.1093/gbe/evw157

Stoltzfus, A., Logsdon, J. M., Palmer, J. D., & Doolittle, W. F. (1997). Intron "sliding" and the diversity of intron positions. *Proceedings of the National Academy of Sciences*, *94*(20), 10739–10744. https://doi.org/10.1073/pnas.94.20.10739

Sverdlov, A. V., Rogozin, I. B., Babenko, V. N., & Koonin, E. V. (2004). Reconstruction of Ancestral Protoempalme Sites. *Current Biology*, *14*(16), 1505–1508. https://doi.org/10.1016/j.cub.2004.08.027

Tarrío, R., Ayala, F. J., & Rodríguez-Trelles, F. (2008). Alternative splicing: A missing piece in the puzzle of intron gain. *Proceedings of the National Academy of Sciences*, *105*(20), 7223–7228. https://doi.org/10.1073/pnas.0802941105

Vinogradov, A. E. (2006). "Genome design" model: Evidence from conserved intronic sequence in human–mouse comparison. *Genome Research*, *16*(3), 347–354. https://doi.org/10.1101/gr.4318206

Wenger, M., & Mathonet, H. (2002). Gendoc: A flexible software documentation generator. *Astronomical Data Analysis Software and Systems XI, 281*, 462–465.