

**EVALUACIÓN DEL PROGRAMA COMPUTACIONAL GEINEXORF-V1 PARA LA  
DETERMINACIÓN DE LA POSICIÓN, FASE DE INTRONES Y TRADUCCIÓN  
DE EXONES OBTENIDOS DE BASES DE DATOS PÚBLICAS.**

**NUBIA ESTHER CASTRO LÓPEZ**

**UNIVERSIDAD INDUSTRIAL DE SANTANDER  
FACULTAD DE CIENCIAS BÁSICAS  
ESCUELA DE BIOLOGÍA  
BUCARAMANGA**

**2016**

**EVALUACIÓN DEL PROGRAMA COMPUTACIONAL GEINEXORF-V1 PARA LA  
DETERMINACIÓN DE LA POSICIÓN, FASE DE INTRONES Y TRADUCCIÓN  
DE EXONES OBTENIDOS DE BASES DE DATOS PÚBLICAS.**

**NUBIA ESTHER CASTRO LÓPEZ**

**Trabajo de Grado presentado como requisito para optar al título de Bióloga**

**Director**

**Francisco José Martínez Pérez**

**Biólogo, Ph.D.**

**UNIVERSIDAD INDUSTRIAL DE SANTANDER**

**FACULTAD DE CIENCIAS BÁSICAS**

**ESCUELA DE BIOLOGÍA**

**BUCARAMANGA**

**2016**

## AGRADECIMIENTOS

A mi mamá, Nubia María López Ramos,  
uno de los mejores seres humanos que he conocido en mi vida.

Al profesor Francisco Martínez,  
por la pasantía y la auxiliatura de investigación.

Y a todas las personas que de una u otra manera  
me colaboraron y permitieron  
llegar al final de este objetivo académico.

## CONTENIDO

INTRODUCCIÓN.....	12
1.OBJETIVOS.....	15
1.1 OBJETIVO GENERAL.....	15
1.2 OBJETIVOS ESPECÍFICOS.....	15
2. MATERIALES Y MÉTODOS.....	16
2.1 - Elaboración y depuración de la base de datos.....	16
2.3 - Validación y funcionamiento del programa computacional GeInExORF-v1...17	17
3.RESULTADOS.....	20
3.1 – BASE DE DATOS DE GENES, mRNA y ORF FUNCIONALES PARA LA VALIDACIÓN DEL GeInExORF_v1 .....	20
3.2 – RESULTADOS OBTENIDOS CON GeInExORF-v1.....	23
3.2.1 - Posición, tamaño de exones e intrones y su fase.....	23
3.2.2 - Traducción de exones con GeInExORF-v1.....	24
3.3 - IDENTIFICACIÓN DE LOS LIMITES DEL FUNCIONAMIENTO DE GeInExORF-v1.....	26
3.3.1 – Extremo 5´ muy largo antes del codón de inicio de la traducción.....	26
3.3.2- Polimorfismo, deleción en el gen, ausencia de parámetros en los algoritmos del GeInExORF-v1 para discriminar la cantidad en pares de bases que corresponden a exones e intrones con relevancia biológica y artificio bioinformático.....	27

3.4 – COMPARACIÓN Y VERIFICACIÓN DE LOS RESULTADOS POSITIVOS A PARTIR DE GelnExORF-v1 CON LOS OBTENIDOS EN SOFTWARES BIOINFORMÁTICOS PÚBLICOS Y POR EL MÉTODO MANUAL.....	29
4. DISCUSIÓN.....	31
5. CONCLUSIONES.....	33
BIBLIOGRAFÍA.....	34

## LISTA DE TABLAS

- Tabla 1. Número de acceso (gi) y la posición para el ORF y el NM para el mRNA, dentro del cromosoma correspondiente, de los genes que codifican para el Receptor de dopamina en *H. sapiens*; correspondiente a las 12 variantes..... 21
- Tabla 2. Número de acceso (gi) y la posición para el ORF y el NM para el mRNA, dentro del cromosoma correspondiente, de los genes que codifican para el Receptor de dopamina en *D. melanogaster*; correspondiente a las 8 variantes..... 22
- Tabla 3. Número de acceso (gi) y la posición para el ORF y el NM para el mRNA, dentro del cromosoma correspondiente, de los genes que codifican para Calcium/calmodulin-dependent protein kinase type II (CAMK2D) en *B. taurus*.....22

## LISTA DE FIGURAS

Figura 1.	Ventana única del software GelnExORF-v1.....	17
Figura 2.	Ventana de dialogo de GelnExORF-v1 cuando el programa arroja resultados.....	18
Figura 3.	Ventana de dialogo de GelnExORF-v1 cuando el programa no arroja resultados.....	19
Figura 4.	Resultados obtenidos con GelnExORF-v1 que muestra la cantidad de exones e intrones y la fase en DR3 VA de <i>H. sapiens</i> .....	24
Figura 5.	Resultado obtenidos con GelnExORF-v1 que muestra el exón 1, los intrones contiguos y los aminoácidos correspondientes a cada codón en DR3 VA de <i>H. sapiens</i> .....	25
Figura 6.	Alineamiento del Gen, mRNA y ORF para DR1 Variante C en el cromosoma 3R en <i>D. melanogaster</i> que muestra un extremo 5' muy largo antes del inicio de la traducción.....	26
Figura 7.	Resultados arrojados por GelnExORF-v1 del análisis de las secuencias de DR3 VX1 de <i>H. sapiens</i> . Los círculos rojos indican los falsos intrones, los azules los polimorfismos, y los blancos las deleciones.....	28
Figura 8.	Resultados y fragmento del alineamiento obtenidos a partir de GelnExORF-v1, del mRNA y la duplicación del ORF que codifican para la CAMK2D en <i>B. taurus</i> .....	28
Figura 9.	Traducción manual y conceptual del exón 1 del gen DR3 variante A, que codifica para el receptor de dopamina en <i>H. sapiens</i> , basado en la traducción conceptual de codón por codón ofrecida por programa bioinformático "Translator" .....	30

## RESUMEN

### TÍTULO

**EVALUACIÓN DEL PROGRAMA COMPUTACIONAL GeInExORF-v1 PARA LA DETERMINACIÓN DE LA POSICIÓN, FASE DE INTRONES Y TRADUCCIÓN DE EXONES OBTENIDOS DE BASES DE DATOS PÚBLICAS.\***

### AUTOR

CASTRO LÓPEZ, Nubia Esther\*\*

### PALABRAS CLAVES

Gen, mRNA, ORF, Intrón, Exón, Traducción, Alineamiento.

### DESCRIPCIÓN

La bioinformática es crucial en el estudio del ensamble de genomas, al aumentar la eficiencia en la identificación de las regiones que componen a los genomas virales, procariontes y eucariontes. Sin embargo, los programas computacionales actuales, diseñados para eucariotas, no consideran la fase del intrón y la traducción de los exones. El método bioinformático de tercera generación GeInExORF-v1 permite la identificación de intrones y exones presentes en una secuencia de ácidos nucleicos. Además, a diferencia de otros, genera el posible marco abierto de lectura de la secuencia de los exones, delimita los intrones y determina su fase. No obstante, dicha herramienta informática no ha sido evaluada por expertos en ciencias naturales. Por lo tanto, se hizo fundamental validar su funcionamiento en genes con diferentes cantidades de intrones, obtenidos de bases de datos públicas como el GenBank. Para ello, se alineó cada Gen, con su respectivo mRNA y ORF, después por medio de GeInExORF-v1 se obtuvieron; los posibles intrones, su fase, los exones con los productos proteicos que codifican y el proceso fisiológico en que pueden participar y se demostró su importancia biológica. Por lo anterior, GeInExORF-v1 facilitará la identificación de nuevas regiones genómicas de forma rápida y eficaz.

---

\* Trabajo de grado.

\*\* Facultad de Ciencias. Escuela de Biología. Director: Francisco José Martínez Pérez, Ph.D.

## ABSTRACT

### TITTLE

**EVALUATION GEINEXORF-V1 COMPUTER PROGRAM FOR DETERMINING THE POSITION OF INTRONS AND TRANSLATION PHASE OF EXONS OBTAINED FROM PUBLIC DATABASES.\***

### AUTHOR

CASTRO LÓPEZ, Nubia Esther\*\*

### KEYWORDS

Gen, RNAm, ORF, Intron, Exon, Translation, Alignment.

### DESCRIPTION

Bioinformatics is a crucial element in the study of genomes assembly, in order to increase efficiency to calculate the identity of nucleotide regions in genomes from virus, prokaryotic and eukaryotic organisms. However, current computer software tools designed to eukaryotic genes do not consider the intron phase and exons translation. The third generation Bioinformatics software GeInExORF-v1 allows identification of exons and introns in a nucleic acids sequence. Furthermore, unlike others, generates a putative open reading frame from exons sequences, delimits introns, and determines their phase. However, experts in natural sciences have not evaluated this software. Consequently, it became essential validate their operation in genes with different amounts of introns obtained from public databases such as GenBank. Therefore, each gene was aligned with their corresponding mRNA and ORF, then GeInExORF-v1 was applied to obtain; the putative introns, its phase, the exons with the encoding protein products, and the physiological process in which they could participate, and was demonstrated their biological importance. For all above GeInExORF-v1 facilitates the identification of new genomic regions quickly and effectively.

---

\* Degree work.

\*\* Science Faculty. Department of Biology. Director: Francisco José Martínez Pérez, Ph.D.

## INTRODUCCIÓN

En la actualidad hay un ritmo cada vez mayor de datos bioinformáticos puestos a disposición, en cuanto a variación genética, transcriptomas, exomas y proteomas. Del mismo modo, una creciente variedad de programas bioinformáticos disponibles a partir de muchas fuentes diversas, que han sido diseñados con la finalidad de identificar un gran número de patrones de secuencias, que tienen una potencial importancia biológica dentro de regiones inter-génicas, genes, transcritos y proteínas (Teber et.al., 2006). Ejemplo de ello; el programa FEX de la plataforma Softberry que determina los extremos 5' y 3' de los posibles exones (Softberry, 2015), las herramientas computacionales para establecer intrones y exones disponible en la página WEB Gene Afinity (Gene Infinity, 2015 [http://www.geneinfinity.org/sp/sp\\_coding.html#splicing](http://www.geneinfinity.org/sp/sp_coding.html#splicing)), la traducción del exón ha sido establecida únicamente por el programa Djinn Lite (Teber et.al., 2006) además, existen otros programas computacionales reportados en la literatura (Goodswen et al., 2012; Jian et al., 2014 y Li et al., 2012).

Sin embargo, los programas computacionales actuales disponibles en el mercado y/o públicos en páginas de internet no consideran la fase del intrón. Para darle una solución a esta problemática el Laboratorio de Genómica de Celomados de la Escuela de Biología de la Universidad Industrial de Santander, ha desarrollado un programa computacional, GeInExORF-v1, que identifica cada intrón con su fase, la correspondiente traducción de cada exón, y la cantidad, posición, tamaño de estos por medio del análisis del alineamiento de las tres secuencias necesarias para codificar una proteína a saber; el Gen, el mRNA y el Marco abierto de lectura (ORF, por sus siglas en Ingles), permitiendo evaluar todas las variantes que codifican para una proteína, siendo esta una de las novedades más importantes ya que se estima que el 75% de eventos de empalme alternativo cambian la secuencia de codificación de las proteínas (Zavolan et al., 2003) que a menudo puede producir isoformas de las proteínas con diferentes composiciones, motivos y dominio (Loraine y Helt, 2002) lo que hace, imprescindible el desarrollo de herramientas bioinformáticas que

permitan modelar las variantes de transcripción. Por lo tanto dentro de este proyecto de grado, se evaluará el funcionamiento y las limitaciones del GelExORF.v1 y sobre todo la relevancia biológica de los resultados, dado que se acumulan secuencias a gran velocidad, se requiere que los biólogos puedan asimilar la información de fuentes diversas relacionada con la secuenciación (Teber et al., 2006).

Es importante, resaltar que este proyecto de grado modalidad pasantía, está incluido dentro del proyecto “Estudio de la factibilidad del uso preferencial de codones para establecer la concentración óptima de desoxinucleótidos (dNTPs), por medio de la secuenciación de genomas con tecnología de tercera generación”, el cual fue aprobado por COLCIENCIAS mediante la convocatoria 700 Convocatoria a proyectos regionales de investigación aplicada para el desarrollo de pruebas de concepto relacionadas con nuevas tecnologías biológicas, biomédicas o asociadas al uso sustentable de energía – 2014, código de contrato FP44842-382-2015-.

## COMPETENCIAS

- 1 - Generó razonamiento lógico así como la habilidad para el manejo y análisis de los resultados de programas computacionales relacionados a la genómica.
- 2 - Dominó el lenguaje molecular y bioinformático.
- 3 – Estableció la capacidad de trabajo en equipo.
- 4 - Desarrolló soluciones alternativas en cuanto a datos de secuenciación genómica e identificación de organización y estructura de genes eucariontes.
- 5 - Elaboró un informe que permite la comunicación escrita de resultados bioinformáticos.
- 6 - Adquirió actitud ética y espíritu de investigación para poder desempeñarse en el campo laboral de la Bioinformática aplicada al Genómica.

## **1. OBJETIVOS**

### **1.1 - OBJETIVO GENERAL**

- Validar el programa computacional GelExORF-v1 creado por el Laboratorio de Genómica de Celomados para la determinación de la posición, fase de intrones y traducción de exones obtenidos de genes reportados en bases de datos públicas.

### **1.2 - OBJETIVOS ESPECÍFICOS**

- Elaborar una base de datos de genes de organismos eucariontes obtenida a partir de una base de datos pública.

- Determinar los posibles marcos de lecturas abiertos con genes con una cantidad menor o mayor a cinco intrones independientemente de la especie.

- Conocer la eficiencia y limitaciones del programa para encontrar intrones y exones cuyos marcos de lectura correspondan a las proteínas biológicas que codifican.

## 2. MATERIALES Y MÉTODOS

### 2.1 - ELABORACIÓN Y DEPURACIÓN DE LA BASE DE DATOS.

Se realizó una búsqueda de genes que codificaran para proteínas, seleccionados a partir de proyectos de genomas de la Base de datos pública del National Center Biotechnology Information de Estados Unidos de América, NCBI (Tatusova et al., 1999).

Para esto, se identificó la proteína y se determinaron los codones que la codifican, lo que corresponde al ORF, el cual debía contener el primer codón ATG que codifica para la primera metionina y los codones de terminación UAG, UGA y UGG para genes Eucariontes (Martínez et al., 2011).

Seguido, se procedió a realizar la búsqueda del mRNA y el ORF en la base de datos, al colocar el nombre de una proteína o de un gen. No se incluyeron mRNA mitocondrial ni cloroplastídico ya que el código genético es diferente al citoplásmico y GelnExOFF-v1 no fue diseñado para proteínas de organelos.

Identificada las secuencias, se generó un formato Fasta para cada Gen, con su respectivo mRNA y ORF. Los cuales se alinearon entre sí por medio del programa computacional Clustal W (Larkin et al., 2007) con los parámetros del modelo evolutivo de genes neuroendocrinos denominado “Pérdida de DNA” (Martínez et al., 2011).

Posteriormente, se analizó el resultado, observando en el alineamiento que las tres secuencias presentaran regiones homologas y no homologas, las cuales correspondían respectivamente a exones o intrones.

Por último, se realizó una revisión bibliográfica para determinar, que la proteína había sido reportada en la base de datos, si se secuenció por métodos experimentales o por algoritmos bioinformáticos (Martínez et al., 2007, 2011).

La base de datos se conformó con genes, que presentaron una cantidad mayor y/o menor de 5 intrones.

## 2.2 – VALIDACIÓN Y FUNCIONAMIENTO DEL PROGRAMA COMPUTACIONAL GEINEXORF-V1.

Puesto que el objetivo de la pasantía es el análisis de los resultados arrojados por el software GeInExORF-v1, mas no el diseño, análisis bioinformático y biomatemático de los algoritmos que lo generaron, a continuación se indican los elementos y métodos requeridos para la obtención de los resultados a analizar.

**Figura 1.** Ventana única del software GeInExORF-v1.



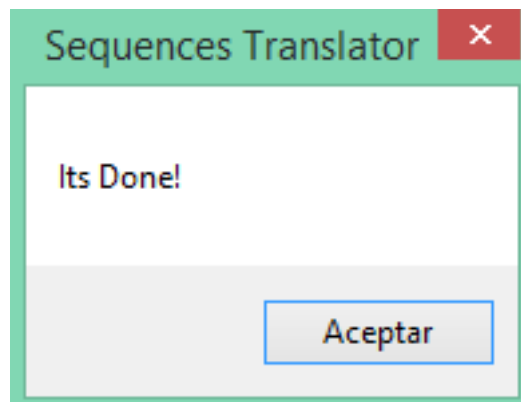
Una vez se activa el programa se observa en la pantalla una ventana única, la cual permitió, el ingreso de la información a analizar. Ella muestra 2 pestañas, “Open” que permitió traer el archivo Fasta que contiene las tres secuencias del Gen, mRNA y el ORF, previamente alineadas, de la carpeta del escritorio del computador o de una memoria USB. Mientras que la pestaña Translate lo ejecuta.

Las 4 subventanas, corresponden a comandos que el usuario no manipula y una de ellas presenta movimiento cuando el programa está corriendo (Fig.1).

Concluido el proceso, GeInExORF-v1 tiene dos opciones, por medio de ventanas, para indicar si las secuencias son o no aptas para ser analizadas. Una de ellas es una ventana de dialogo, que indica que el archivo es correcto y generó resultados (Fig. 2), mientras la otra explica la razón por la que no arrojó los resultados y da una serie de instrucciones para revisar las secuencias (Fig. 3).

Los resultados se encuentran disponibles en la direccion de salida que se indica en la subventana Final File (Fig.1) y por supuesto el formato Fasta debe cumplir con las características impuestas por el diseñador, para que el GeInExORF-v1 lo lea correctamente.

**Figura 2.** Ventana de dialogo de GeInExORF-v1 cuando el programa arroja resultados.



**Figura 3.** Ventana de dialogo de GeInExORF-v1 cuando el programa no arroja resultados.



### 3. RESULTADOS

#### 3.1 – BASE DE DATOS DE GENES, mRNA y ORF FUNCIONALES PARA LA VALIDACIÓN DEL GeInExORF-v1.

Para la construcción de la base de datos y sustentados en los criterios de selección, expuestos en la metodología, del gen para la validación del GeInExORF-v1, en especial la cantidad de intrones contenidos en el ORF, se eligió a la súperfamilia de receptores de dopamina y al gen proteína quinasa tipo II dependiente de calcio/calmodulina (CAMKA2D) correspondientes a las siguientes especies:

El primero corresponde a los genes que codifican para la súper familia de los receptores de dopamina en *Homo sapiens*, se encontraron los 5 genes que la constituyen. DR1 y DR5 sin intrones y Dr2-Dr4 con más de 2 intrones. Además, las 12 variantes de mRNA con sus respectivos ORF (Tabla 1).

También, se incluyó a los dos genes que codifican para el receptor antes mencionado en *Drosophila melanogaster* del cual se encontraron las cinco variantes de mRNA para el DR1 y tres para el DR2; y la misma cantidad de ORF (Tabla 2).

En tanto que, para el gen que codifica para la CAMK2D se seleccionó a *Bos taurus*, ya que este gen tiene más de dos intrones dentro de 307433 pares de bases (pb), es decir 4.35 veces más grande que el DR3 el gen más extenso de dopamina (Tabla 3).

Se encontró que los genes obtenidos son producto de procesos bioinformáticos, es decir inteligencia artificial, mientras que los mRNA y ORF son productos experimentales como ha sido reportado en el proyecto de secuenciación del genoma humano (Kidd et al. 2008).

**Tabla 1.** Número de acceso (gi) y la posición para el ORF y el NM para el mRNA, dentro del cromosoma correspondiente, de los genes que codifican para el Receptor de dopamina en *H. sapiens*; correspondiente a las 12 variantes.

ESPECIE	<i>Homo sapiens</i>	
PROTEINA	Receptor de Dopamina	
SECUENCIAS	gi	POSICIÓN ó No. ACCESO
DR1 GEN CROMOSOMA 5	226509980	5001-8489
DR1 RNAm	88758587	NM_000794.3
DR1 ORF	88758587	946-2286
DR2 GEN CROMOSOMA 11	209977039	5001-70685
DR2 V1 RNAm	181336946	NM_000795.3
DR2 V1 ORF	181336946	236-1567
DR2 V2 RNAm	181337009	NM_016574.3
DR2 V2 ORF	181337009	236-1480
DR3 GEN CROMOSOMA 3	553727254	5001-75756
DR3 VA RNAm	595763542	NM_000796.5
DR3 VA ORF	595763542	432-1634
DR3 VE RNAm	595763546	NM_033663.5
DR3 VE ORF	595763546	432-1535
DR3 VF RNAm	595763545	NM_001282563.2
DR3 VF ORF	595763545	291-1493
DR3 VG RNAm	595763543	NM_001290809.1
DR3 VG ORF	595763543	424-1626
DR3 X1 RNAm	767925935	XM_011512510.1
DR3 X1 ORF	767925935	513-1715
DR3 X2 RNAm	767925937	XM_011512511.1
DR3 X2 ORF	767925937	421-1623
DR3 X3 RNAm	767925939	XM_011512512.1
DR3 X3 ORF	767925939	628-1830
DR4 GEN CROMOSOMA 11	568815425	636234-639634
DR4 RNAm	296010830	NM_000797.3
DR4 ORF	296010830	1-1260
DR5 GEN CROMOSOMA 4	237681089	5001-7376
DR5 RNAm	171919773	NM_000798.4
DR5 ORF	171919773	397-1830

**Tabla 2.** Número de acceso (gi) y la posición para el ORF y el NM para el mRNA, dentro del cromosoma correspondiente, de los genes que codifican para el Receptor de dopamina en *D. melanogaster*; correspondiente a las 8 variantes.

ESPECIE	<i>Drosophila melanogaster</i>	
PROTEINA	RECEPTOR DE DOPAMINA	
SECUENCIA	gi	POSICIÓN ó NM
DR1 Cromosoma 3R	671162122	14214810-14165379
DR1 VB RNAm	386765733	NM_057659.4
DR1 VB ORF	386765733	534-2024
DR1 VC RNAm	386765734	NM_001170136.2
DR1 VC ORF	386765734	534-1979
DR1 VD RNAm	386765735	NM_001260163.1
DR1 VD ORF	386765735	489-202
DR1 VE RNAm	442619031	NM_001275634.1
DR1 VE ORF	442619031	534-2216
DR1 VF RNAm	939620353	NM_001316525.1
DR1 VF ORF	939620353	534-2141
DR2 Cromosoma 3R	671162122	29659984-29630304
DR2 VA RNAm	281362799	NM_170420.2
DR2 VA ORF	281362799	904-2517
DR2 VB RNAm	665395361	NM_079824.4
DR2 VB ORF	665395361	904-2523
DR2 VC RNAm	442621700	NM_001276143.1
DR2 VC ORF	442621700	904-3327

**Tabla 3.** Número de acceso (gi) y la posición para el ORF y el NM para el mRNA, dentro del cromosoma correspondiente, de los genes que codifican para Calcium/calmodulin-dependent protein kinase type II (CAMK2D) en *B. taurus*.

ESPECIE	<i>Bos taurus</i>	
GEN	CAMK2D	
SECUENCIA	gi	POSICIÓN
CAMK2D Cromosoma 3R	957816011	13105861-13413294
CAMK2D RNAm	114052473	NM_0001046333.1
CAMK2D ORF	114052473	39-1505

## 3.2 –RESULTADOS OBTENIDOS CON GelnExORF-v1

Los resultados obtenidos de la validación de GelnExORF-v1, con la base de datos construida en el punto 3.1, se presentaran por secciones para que se tenga un mejor análisis y comprensión de estos.

**3.2.1 - Posición, tamaño de exones e intrones y su fase.** El programa arroja resultados con información que corresponde a los parámetros bioinformáticos de la posición y tamaño en pares de bases de lo que puede ser considerado un intrón y/o un exón.

Para la súperfamilia de los receptores para la dopaminas el GelnExORF-v1 arrojó resultados para 10 de los 12 mRNA en *H. sapiens* y para 2 de los 8 mRNA en *D. melanogaster*.

La figura 4 muestra de manera explícita la información que se obtiene con el GelnExORF-v1 sobre la posición, la longitud, y la fase de cada uno de los intrones y la cantidad de exones e intrones. Este ejemplo contiene los resultados que se obtuvieron para el gen DR3 Variante A con su respectivo mRNA y ORF, que codifica para dopamina *H.sapiens*, datos reportados en el GenBank (Tabla 1).

**Figura 4.** Resultados obtenidos con GeInExORF-v1 que muestra la cantidad de exones e intrones y la fase en DR3 VA de *H. sapiens*.

```
Dr3 Var A Align2.aln - Result.fox
1 *****
2 ***** Introns and Exons *****
3 *****
4
5 -----
6 Document: Dr3 Var A Align2.aln.txt
7 -----
8
9 ***** Introns *****
10
11 Posicion = 0          Longitud = 20355      Fase = X
12 Posicion = 20752     Longitud = 6629       Fase = 0
13 Posicion = 27689     Longitud = 11855      Fase = 1
14 Posicion = 39656     Longitud = 12197      Fase = 1
15 Posicion = 51994     Longitud = 7718       Fase = 0
16 Posicion = 59908     Longitud = 8099       Fase = 0
17 Posicion = 68291     Longitud = 2205       Fase = 1
18
19 Total Introns: 7
20
21
22 ***** Exons *****
23
24 Posicion = 27415     Longitud = 274
25 Posicion = 39544     Longitud = 112
26 Posicion = 51853     Longitud = 141
27 Posicion = 59712     Longitud = 196
28 Posicion = 68007     Longitud = 284
29 Posicion = 70496     Longitud = 196
30
31 Total Exons: 6
32
33 -----
34
35
```

**3.2.2 - Traducción de exones con GeInExORF-v1.** GeInExORF-v1 utiliza el término "Traslation" para describir las posibles traducciones de las regiones de transcripción básica, tales como exones.

El formato de cómo se muestran los resultados permite una visión general del tamaño físico del mapa de genes y sus transcripciones asociadas, en términos de los tamaños relativos de los intrones, exones, y la densidad de características a lo largo de estos mapas. Cada aminoácido es colocado en la parte superior del segundo nucleótido de cada condón y ambos se encuentran acorde a la nomenclatura propuesta por el Comité Internacional de la Unión de Bioquímica (NC-IUB, 1986).

Es de llamar la atención que muchos de los nucleótidos tiene diferente color entre ellos, pero los desarrolladores del software no suministraron información del motivo de esta presentación. Pese a lo anterior, la visualización del exón con respecto al intrón hace sencillo discernir entre uno y el otro. Prueba de ello es la imagen fotográfica que muestra el alineamiento de las tres secuencias Gen, RNAm, ORF y la traducción conceptual a aminoácidos de los codones que componen al exón 1 del gen DR3 Variante A de *H. sapiens*. (Fig 5).

**Figura 5.** Resultado obtenidos con GeInExORF-v1 que muestra el exón 1, los intrones contiguos y los aminoácidos correspondientes a cada codón en DR3 VA de *H. sapiens*.

```

Dr3 Var A Align2.aln - Result.fox
2843 Alignment
2844
2845 Translation ----- T A V V M P V H Y Q H G T G Q S S C R R V A L M I T
2846 GeneH.Chromosome3 CTTGCTTTTCAGGTACACTGCAGTGGTCATGCCCGTTCACTACCAGCATGGCACGGGACAGAGCTCCTGTGCGGCGGTGGCCCTCATGATCACGG
2847 mRNAH.Chromosome3 -----CACTGCAGTGGTCATGCCCGTTCACTACCAGCATGGCACGGGACAGAGCTCCTGTGCGGCGGTGGCCCTCATGATCACGG
2848 ORFH.Chromosome3 -----CACTGCAGTGGTCATGCCCGTTCACTACCAGCATGGCACGGGACAGAGCTCCTGTGCGGCGGTGGCCCTCATGATCACGG
2849 Alignment *****
2850
2851 Translation A F A V S C P L L F G F N T T G-----
2852 GeneH.Chromosome3 CCTTTGCTGTGCTCCTGCCCTCTTCTGTTGGCTTTAATACCACAGGTAACAGTGATGGCTTCATTTCTAGCATCCATGTGATTCACAGCACAGCTG
2853 mRNAH.Chromosome3 CCTTTGCTGTGCTCCTGCCCTCTTCTGTTGGCTTTAATACCACAGG-----
2854 ORFH.Chromosome3 CCTTTGCTGTGCTCCTGCCCTCTTCTGTTGGCTTTAATACCACAGG-----
2855 Alignment *****
2856
2857 Translation -----
2858 GeneH.Chromosome3 GTTGTGCTCAGCCCTGCAATTTCTGACCCCGGTCAGTGTACTTTGCAATTAAATGAGAATGCCACAGAATGCAAGCTTTGGCCTTTGTGGGTTA
2859 mRNAH.Chromosome3 -----
2860 ORFH.Chromosome3 -----
2861 Alignment
2862

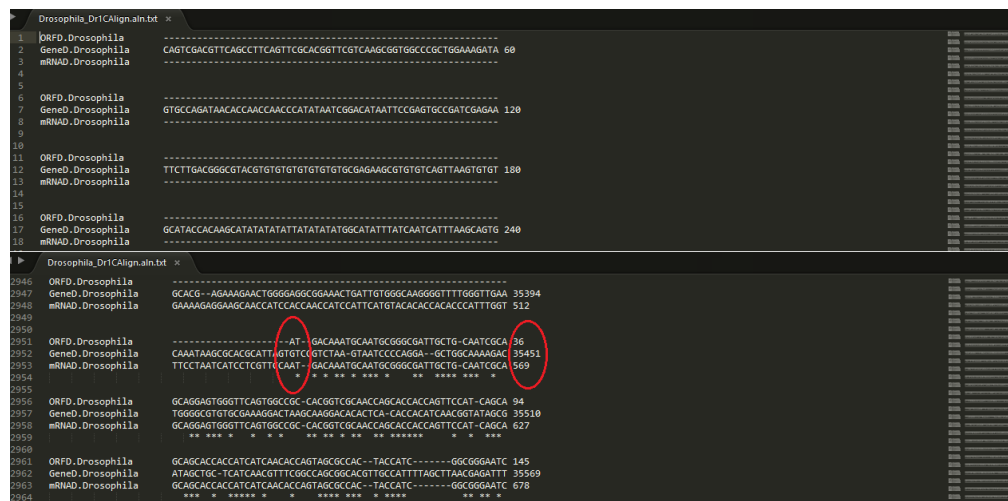
```

### 3.3 – IDENTIFICACIÓN DE LOS LIMITES DEL FUNCIONAMIENTO DE GelnExORF-v1.

Se determinaron diferentes elementos que generaron dos opciones que cuestionaron la veracidad de los datos obtenidos en los resultados. Siendo estas; la no ejecución del programa y la segunda; la generación de falsos positivos. A continuación, se detallan.

**3.3.1 - Extremo 5' muy largo antes del codón de inicio de la traducción.** Para este caso, no se obtuvo un resultado positivo debido a que desde un principio el primer exón que contiene el codón ATG que indica el sitio de inicio de la traducción, no fue identificado desde el alineamiento del gen. Razón de ello, se observaron alineamientos inespecíficos con respecto a otras partes del gen. Lo anterior fue debido al número de pares de bases del sitio de inicio entre el primer exón y del segundo exón. Ejemplo de ello; se muestra en la Figura 6 donde se observa que el extremo 5' es muy largo en el alineamiento del Gen, que codifica para el receptor de Dopamina, DR1 Variante C en el cromosoma 3R en *D. melanogaster*.

**Figura 6.** Alineamiento del Gen, mRNA y ORF para DR1 Variante C en el cromosoma 3R en *D. melanogaster* que muestra un extremo muy largo antes del inicio de la traducción.



**3.3.2- Polimorfismo, delección en el gen, ausencia de parámetros en los algoritmos del GelnExORF-v1 para discriminar la cantidad en pares de bases que corresponden a exones e intrones con relevancia biológica y arteficio bioinformático.** Cuando se presentaba en el alineamiento un espacio de 2 o más nucleótidos, GelnExORF-v1 consideraba esta incoherencia biológica como un intrón y por lo tanto arrojaba resultados. El patrón anterior se presentó en alineamientos con polimorfismos y/o una o más delecciones entre las secuencias. Ejemplo de los tres casos se muestran en la figura 7 con los resultados obtenidos del análisis de las secuencias Gen, mRNA y ORF correspondientes al receptor de dopamina DR3 variante X1 del cromosoma 3 en *H. sapiens*.

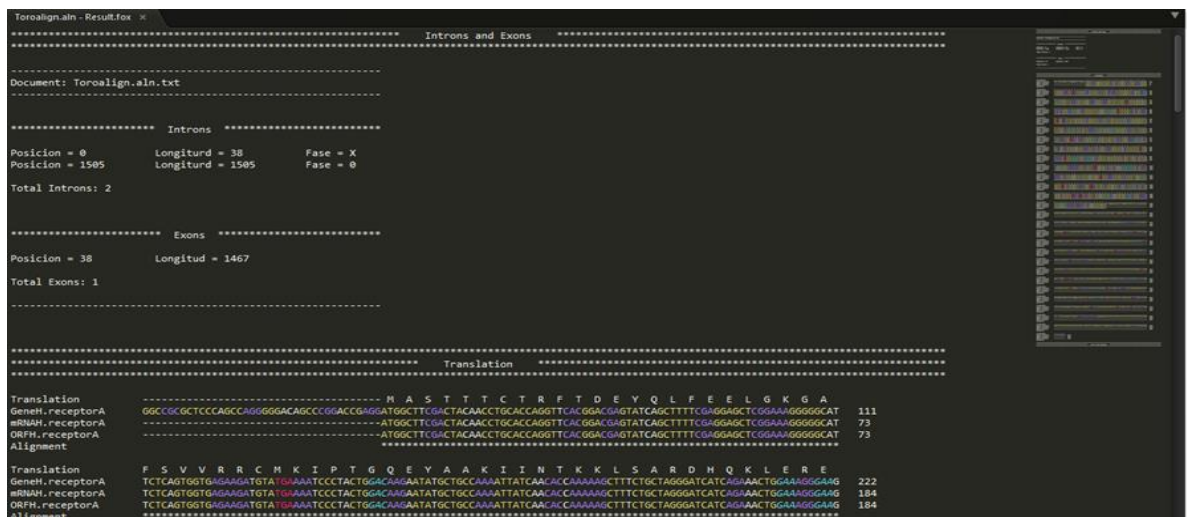
Así mismo, un aspecto sobresaliente fue que GelnExORF-v1 no podía ser utilizado si Clustal W 2.1 no generaba alineamientos por diferencias significativas entre la cantidad de nucleótidos del gen y las otras dos secuencias. Por ello, se generó un arteficio bioinformático, al alterar el formato Fasta para lograr el alineamiento, mediante el intercambio de la secuencia original del gen por la del mRNA, la del mRNA por la del ORF, la cual fue constante, es decir el formato contenía dos secuencias idénticas del ORF. GelnExORF-v1 no tuvo la capacidad de discernir e identificar el error de la ausencia de una de las secuencias, es decir la del gen, y aun así arrojó resultados. Cabe aclarar que los nombres se mantuvieron para que el formato fuese leído por el programa.

Todo lo anterior se observa en la figura 8 que muestra los resultados obtenidos de la alteración de las secuencias que codifican para CAMCK2 en *B. taurus*, cuyo gen tiene un tamaño 307433 pb en comparación con los 3010 pb del mRNA y 1467 pb del ORF.

**Figura 7.** Resultados arrojados por GeInExORF-v1 del análisis de las secuencias de DR3 VX1 de *H. sapiens*. Los círculos rojos indican los falsos intrones, los azules los polimorfismos, y los blancos las deleciones.



**Figura 8.** Resultados y fragmento del alineamiento obtenidos a partir de GeInExORF-v1, del mRNA y la duplicación del ORF que codifican para la CAMK2D en *B. taurus*.



### **3.4 – COMPARACIÓN Y VERIFICACIÓN DE LOS RESULTADOS POSITIVOS A PARTIR DE GelnExORF-v1 CON LOS OBTENIDOS EN SOFTWARES BIOINFORMÁTICOS PÚBLICOS Y POR EL MÉTODO MANUAL.**

Con el fin de comparar las formas en que se observan los resultados de GelnExORF-v1 *versus* programas en línea existentes, a saber GeneSplicer (Pertea et al., 2001), SplicerPort (Dogan et al., 2007), Splice Site Prediction by Neural Network (BDGP), GeneMark (Besemer y Borodovsky 2005), y Alternative Splice Site Predictor (ASSP) (Wang y Marín 2006) se replicó el análisis de la secuencias para el receptor de Dopamina DR3 VA de *H. sapiens* (Anexo 1). Se determinó que la forma y estilo no era útil y no tenía una importancia biológica, puesto que mostraba una cantidad ilógica de intrones y/o de la secuencias consensos que los define. Estos están basados en algoritmos intrínsecos de la página WEB.

Finalmente, se hizo necesario corroborar la relevancia biológica de los datos obtenidos con GelnExORF-v1, de forma manual con la identificación de los codones y su respectivo aminoácido como se ejemplifica en la figura 10, para el gen mencionado en esta sección. En ella se muestra la concordancia entre el exón, su traducción, el intrón inmediato y el primer nucleótido del codón que determinara la fase con el siguiente exón.

**Figura 9.** Traducción manual y conceptual del exón 1 del gen DR3 variante A, que codifica para el receptor de dopamina en *H. sapiens*, basado en la traducción conceptual de codón por codón ofrecida por programa bioinformático “Translate”.

```

                                     M A S L S Q L
gi189191860_432-1634 -----ATGGCATCTCTGAGCCAGCT 20
GeneH_sapiens CACAGGAAAGCCCTTGGCATCACGCACCTCCCTCTGGGCTATGGCATCTCTGAGCCAGCT 7080
mRNAH_sapiensD3 -----AAGCCCTTGGCATCACGCACCTCCCTCTGGGCTATGGCATCTCTGAGCCAGCT 451
                                     ..

S G H L N Y T C G A E N S T G A S Q A R
gi189191860_432-1634 GAGTGGCCACCTGAACTACACCTGTGGGGCAGAGAACTCCACAGGTGCCAGCCAGGCCCG 80
GeneH_sapiens GAGTGGCCACCTGAACTACACCTGTGGGGCAGAGAACTCCACAGGTGCCAGCCAGGCCCG 7140
mRNAH_sapiensD3 GAGTGGCCACCTGAACTACACCTGTGGGGCAGAGAACTCCACAGGTGCCAGCCAGGCCCG 511
                                     ..

P H A Y Y A L S Y C A L I L A I V F G N
gi189191860_432-1634 CCCACATGCCTACTATGCCCTCTCCTACTGCGCGCTCATCCTGGCCATCGTCTTCGGCAA 140
GeneH_sapiens CCCACATGCCTACTATGCCCTCTCCTACTGCGCGCTCATCCTGGCCATCGTCTTCGGCAA 7200
mRNAH_sapiensD3 CCCACATGCCTACTATGCCCTCTCCTACTGCGCGCTCATCCTGGCCATCGTCTTCGGCAA 571
                                     ..

G L V C M A V L K E R A L Q T T T N Y L
gi189191860_432-1634 TGGCCTGGTGTGCATGGCTGTGCTGAAGGAGCGGGCCCTGCAGACTACCACCAACTACTT 200
GeneH_sapiens TGGCCTGGTGTGCATGGCTGTGCTGAAGGAGCGGGCCCTGCAGACTACCACCAACTACTT 7260
mRNAH_sapiensD3 TGGCCTGGTGTGCATGGCTGTGCTGAAGGAGCGGGCCCTGCAGACTACCACCAACTACTT 631
                                     ..

V V S L A V A D L L V A T L V M P W V V
gi189191860_432-1634 AGTAGTGAGCCTGGCTGTGGCAGACTTGCTGGTGGCCACCTTGGTGATGCCCTGGGTGGT 260
GeneH_sapiens AGTAGTGAGCCTGGCTGTGGCAGACTTGCTGGTGGCCACCTTGGTGATGCCCTGGGTGGT 7320
mRNAH_sapiensD3 AGTAGTGAGCCTGGCTGTGGCAGACTTGCTGGTGGCCACCTTGGTGATGCCCTGGGTGGT 691
                                     ..

Y L E V
gi189191860_432-1634 ATACCTGGAGGTGA----- 274
GeneH_sapiens ATACCTGGAGGTGAGTAGACTTCAGGTGCATGTTGTCTCTATGACTGTGCTAGTACTTGT 7380
mRNAH_sapiensD3 ATACCTGGAGGTGA----- 705
                                     ..

```

## 4. DISCUSIÓN

El GelNExORF-v1 fue diseñado para facilitar el análisis de la estructuras de intrones y exones en genes eucariontes, por lo que era indispensable validar su funcionamiento y sus resultados desde una visión biológica además de la bioinformática. Los programas de búsqueda comparativos de genes realizan la búsqueda basados únicamente en patrones que definen dónde un gen comienza y termina (International Human Genome Sequencing Consortium, 2004). GelNExORF-v1 funciona mediante la comparación de los segmentos de las secuencias que codifican para proteínas. Además, posee las características necesarias para la visualización de los resultados, tales como la facilidad de navegación, código de colores, y las representaciones a nivel macro y detallado del alineamiento de las tres secuencias (Teber et al., 2006).

La relevancia biológica de los resultados inicialmente se ve comprometida por la construcción de una base de datos con estructuras génicas canónicas para el formato Fasta. Durante este proceso se presentaron considerables desafíos, debido a la densa información que ofrece el GenBank y a que existen numerosas secuencias para el Gen, mRNA, ORF y proteína que poseen deleciones, inserciones, nucleótidos y/o aminoácidos no determinados y a errores e inexactitudes que puedan existir en las principales bases de datos de secuencias (Wesche et al., 2004). Estas inconsistencias en las anotaciones surgen debido a los métodos en la identificación de los genes y en la generación de los ORF (Teber et al., 2006).

No obstante, mediante el alineamiento de las tres secuencias necesarias para codificar una proteína, GelNExORF-v1, determina la posición y la fase de los intrones, como se observa en la figura 4, esto tiene una gran importancia ya que los algoritmos bioinformáticos actuales se centran únicamente en la predicción de las regiones codificantes más no en las regiones no traducidas (Ashurst y Collins,

2003) es decir, que las herramientas bioinformáticas disponibles en línea únicamente se enfocan en la predicción del exón, de las proteínas y de la estructura secundaria (Rutherford et al., 200). Además de esta novedad, permite analizar individualmente las secuencias específicas para cada variante (tabla 1) lo cual es trascendental ya que el procesamiento alternativo del ARN mensajero es un importante mecanismo de regulación de la expresión génica en mamíferos (Stamm et al., 2006) por ejemplo en humanos entre un 30% y un 50% de todos los genes produce múltiples variantes de la transcripción (Lander et al., 2001).

La ventaja de estos métodos es que dan resultados automáticos y rápidos, sin embargo se paga un precio en términos de fiabilidad (Kleffe et al., 1997). Tal cual sucede con GelExORF-v1, el cual no discrimina la cantidad de deleciones encontradas y si estas deleciones en el alineamiento son concordantes entre el mRNA y el ORF, siendo apenas dos nucleótidos los considera como un intron arrojando resultados sin relevancia biológica como se observa en la figura 7.

Además, de que estos modelos bioinformático son específicos para modelos humanos o unos pocos organismos bien estudiados (Kleffe et al., 1997). Consecuentemente, la interpretación de las diferentes secuencias génicas, únicamente por este método computacional es insuficiente para generar resultados satisfactorios para toda la gama de las secuencias reportadas que codifican para proteínas. Por lo tanto se hace necesario coordinar los resultados computacionales con el método manual para aprender de las inconsistencias generadas entre las estructuras de genes (Pennisi, 2003). Por lo tanto, sigue siendo necesaria la traducción conceptual y manual de cada codón con la ayuda de TRANSLATE, a pesar de parecer un método obsoleto por el tiempo que se necesita para llevarlo a cabo.

## 5. CONCLUSIONES

Los resultados obtenidos del programa computacional GelnExORF v-1 tuvieron una relevancia biológica mostrando la fase y la posición de los intrones y la traducción de los exones, concordantes con la teoría, comprobando así que es un programa óptimo para poner en funcionamiento y facilitar parte de la biología molecular. Además de que GelnExORF-v1 es sencillo de manejar y tiene una plataforma para cualquier Windows.

Otra ventaja adicional del funcionamiento de GelnExORF-V1 es debido a que detecta de forma indirecta diferencias entre el gen y el mRNA cuando no corresponden a la fecha y/o sistema de secuenciación sea experimental o *in silico*. Lo que contribuye para la toma de decisiones en la clonación de un segmento genómico o de cDNA de forma experimental.

## BIBLIOGRAFÍA

ARTIMO, Panu. JONNALAGEDDA, Manohar. KONSTANTIN, Arnold. BARATIN, Delphine. CSARDI, Gabo. DE CASTRO, Edouard. DUVAUD, Séverini. FLEGEL, Volker. FORTIER, Arnaud. GASTEIGER, Elisabeth. GROSDIDIER, Aurélien. HERNANDEZ, Céline. IOANNIDIS, Vassilios. KUZNETSOV, Dmitry. LIECHTI, Robin. MORETTI, Sébastien. MOSTAGUIR, Khaled. REDASCHI, Nicole. ROSSIER, Grégori. XENARIOS, Ioannis. STOCKINGER, Heinz. ExPASy. En: SIB bioinformatics resource portal, Nucleic Acids Res. 2012. Vol. 40, No. 1, P. 597-603.

ASHURST, Jennifer. COLLINS, John. Gene annotation: prediction and testing. En: Annu Rev Genomics Hum Genet. 2003. Vol. 4, p. 69-88.

Berkeley Drosophila Genome Project. BDGP: Splice Site Prediction by Neural Network. Available from internet: <[www.fruitfly.org/seq\\_tools/splice.html](http://www.fruitfly.org/seq_tools/splice.html).>

BESEMER, John. BORODOVSKY, Mark. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. En: Nucleic Acids Res. Jul, 2005. Vol. 1, No. 33, p. 451-454.

DOGAN, Rezart. GETOOR, Lise. WILBUR, John. MOUNT, Stephen. SplicePort- an interactive splice-site analysis tool. En: Nucleic Acids Res. Jul, 2007. Vol.35, p. 285-91.

Gene Infinity, 2015. Available from internet:

<[http://www.geneinfinity.org/sp/sp\\_coding.html#splicing](http://www.geneinfinity.org/sp/sp_coding.html#splicing).>

GOODSWEN, Stephen. KENNEDY, Paul. ELLIS, John. Evaluating high-throughput ab initio gene finders to discover proteins encoded in eukaryotic pathogen genomes missed by laboratory techniques. En: PLoS One. 2012. Vol. 7, No.11, p.609.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. En: Nature. 2004, Vol. 431, p. 931-945.

JIAN, Xuequi. BOERWINKLE, Eric. LIU, Xiaoming. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. En: Genet Med. Jul, 2014. Vol. 16, No. 7, p. 497-503.

KIDD, Jeffrey. COOPER, Gregory. DONAHUE, William. HAYDEN, Hilary. SAMPAS, Nick. GRAVES, Tina. HANSEN, Nancy. TEAGUE, Brian. ALKAN, Can. ANTONACCI, Francesca. HAUGEN, Erick. ZERR, Troy. YAMADA, Alice. TSANG, Peter. NEWMAN, Tera. TÜZÜN, Eray. CHENG, Ze. EBLING, Heather. TUSNEEM, Nademm. DAVID, Robert. GILLETT, Will. PHELPS, Karen. WEAVER, Molly. SARANGA, David. BRAND, Adrienne. TAO, Wei. GUSTAFSON, Erick. MCKERNAN, Kevin. CHEN, Lin. MALIG, Maika. SMITH, Joshua. KORN, Joshua. MCCARROLL, Steven. ALTSHULER, David. PEIFFER, Daniel. DORSCHNER, Michael. STAMATOYANNOPOULOS, John. SCHWARTZ, David. NICKERSON,

Deborah. MULLIKIN, James. WILSON, Richard. BRUHN, Laurakay. OLSON, Maynard. KAUL, Rajinder, SMITH, Douglas. EICHLER Evan. Mapping and sequencing of structural variation from eight human genomes. En: Nature. May, 2008. Vol. 453, p. 56-64.

KLEFFE, Jürgen. HERMANN, Klaus. VAHRSON, Wolfgang. WITTIG, Burghardt. BRENDDEL, Volker. GeneGenerator-a flexible algorithm for gene prediction and its application to maize sequences. En: Bioinformatics.1998. Vol. 14, No. 3, p. 232-243.

LANDER, ES et al. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. En: Nature. 2001. Vol. 409, p. 860-921.

LARKIN, MA. BLACKSHIELDS, G. BROWN, NP. CHENNA, R. MCGETTIGAN, PA. MCWILLIAM, H. VALENTIN, F. WALLACE, IM. WILM, A. LOPEZ, R. THOMPSON, JD. GIBSON ,TJ. HIGGINS, DG. Clustal W and ClustalX version 2. En: Bioinformatics. 2007. Vol. 23, No. 21, p. 2947-2948.

LI, Chenhong. RIETHOVEN, Jean-Jack. NAYLOR, Gavin. EvolMarkers: a database for mining exon and intron markers for evolution, ecology and conservation studies. En: Mol Ecol Resour. Sep, 2012. Vol. 12, No. 5, p. 967-71.

LORAINE, Ann. Helt, Gregg. Visualizing the genome: techniques for presenting human genome data and annotations. En: BMC Bioinformatics. 2002. (3):19-26.

MARTÍNEZ-PÉREZ, Francisco. DURÁN-GUTIÉRREZ, Darinka. DELAYE, Luis. BECERRA, Arturo. AGUILAR, Guadalupe. ZINKER, Samuel. Loss of DNA: a plausible molecular level explanation for crustacean neuropeptide gene evolution. En: Peptides. Jan, 2007. Vol. 28, No.1, p.76-82.

MARTÍNEZ-PÉREZ, Francisco. BENDENA, William. CHANG, Belinda. TOBE, Stephen. Influence of codon usage bias on FGLamide-allatostatin mRNA secondary structure. En: Peptides. Mar, 2011. Vol. 32, No. 3, p. 509-17.

NC-IUB. Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Nomenclature Committee of the International Union of Biochemistry (NC-IUB). En: Proc Natl Acad Sci U S A. Jan, 1986. Vol. 83, No. 1, p. 4-8.

PENNISI, Elizabeth. Bioinformatics. Gene counters struggle to get the right answer. En: Science. 2003. Vo. 301, p. 1040-1041.

PERTEA, Mihaela. LIN, Xiaoying. SALZBERG, Steven. GeneSplicer: a new computational method for splice site prediction. En: Nucleic Acids Res. Mar, 2001. Vol. 29, No. 5, p.1185-90.

RUTHERFORD, Kim. PARKHILL, Julian. CROOK, James. HORSNELL, Terry. RICE, Peter. RAJANDREAM, Maria-Adéle. BARRELL, Bart. Artemis: sequence visualization and annotation. En: Bioinformatics. 2000. Vol. 16, p. 944-945.

STAMM, Stefan. BEN-ARI, Shani. RAFALSKA, Ilona. TANG, Yesheng. ZHANG, Zhaiyi. TOIBER, Debra. THANARAJ, T.A, SOREQ, Hermona. Function of alternative splicing. En: Gene. 2005. Vol. 344, p. 1-20.

Softberry programs, 2015. Available from internet:  
<<http://www.softberry.com/berry.phtml?topic=fex&group=programs&subgroup=gfin>>

TATUSOVA, Tatiana, KARSCH-MIZRACHI, Ilene. OSTELL, James. Complete genomes in WWW Entrez: data representation and analysis. En: Bioinformatics. Jul-Aug, 1999. Vol. 15, No. 7, p.536-43.

TEBER, Erdhal. CRAWFORD, Edward. BOLTON, Ken. VAN DYK, Dereck.SCHOFIELD, Peter. KAPOOR, Vimal. CHUECH, Bret. Djinn Lite: a tool for customised gene transcript modelling, annotation-data enrichment and exploration. En: BMC Bioinformatics. Jan, 2006. Vol. 23, p. 7-33.

WANG, Magnus. MARÍN, Antonio. Characterization and Prediction of Alternative Splice Sites. En: Gene. 2006. Vol .366, p. 219-227.

WESCHE, Philipp. GAFFNEY, Daniel. KEIGHTLEY, Daniel. DNA sequence error rates in Genbank records estimated using the mouse genome as a reference. En: DNA Seq. 2004. Vol. 15, p. 362-364.

ZAVOLAN, Michaela. KONDO, Shinji. SCHÖNBACH, Christian. ADACHI, Jun. HUME, David. HAYAZHIZAKI, Yoshihide. Y GAASTERLAND, Terry. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. En: Genome Research. 2003. Vol.13. p. 1290-1300.