COLEGIO DE POSTGRADUADOS

INSTITUCION DE ENSEÑANZA E INVESTIGACION EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

POSTGRADO DE SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA

ESTADÍSTICA

DOS MÉTODOS PARA LA PREDICCIÓN DE INTRONES Y EXONES EN UN GEN

EVELIO HERNÁNDEZ JUÁREZ

T E S I S

PRESENTADA COMO REQUISITO PARCIAL
PARA OBTENER EL GRADO DE:

DOCTOR EN CIENCIAS

MONTECILLO, TEXCOCO, EDO. DE MEXICO

2015

La presente tesis titulada "**DOS MÉTODOS PARA LA PREDICCIÓN DE INTRONES Y EXONES EN UN GEN**" realizada por el alumno: Evelio Hernández Juárez, bajo la dirección del Consejo Particular indicado, ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

DOCTOR EN CIENCIAS

SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA

ESTADÍSTICA

CONSEJO PARTICULAR

CONSEJERO _____

Dr. Gustavo Ramírez Valverde

ASESOR _____

Dr. Sergio Pérez Elizalde

ASESOR _____

Dr. Amalio Santacruz Varela

ASESOR _____

Dr. Benito Ramírez Valverde

ASESOR _____

Dr. David Sotres Ramos

Montecillo, Texcoco, Estado de México, Agosto de 2015

DOS MÉTODOS PARA LA PREDICCIÓN DE INTRONES Y EXONES EN UN GEN

Evelio Hernández Juárez, Dr.

Colegio de postgraduados, 2015

RESUMEN

Para organismos eucariontes, en un gen, además de las regiones codificantes de proteínas (exones) se encuentran regiones no codificantes conocidas como intrones. Existen diferentes enfoques para encontrar las fronteras entre intrones y exones, uno de ellos trata de distinguir cambios en la composición (proporción) de sus nucleótidos y en este contexto usar la segmentación recursiva basada en medidas de divergencia; en esta metodología se necesita definir un "umbral" (parámetro que sirve como criterio de parada). En este estudio se propone que la elección del umbral sea basada en la distribución Monte Carlo de la máxima divergencia, y que dicho umbral dependa del tamaño de la secuencia considerada.

Bajo la misma perspectiva de encontrar diferencias en la proporción de nucleótidos dentro de las secuencias, se propone usar la búsqueda de puntos de cambio en datos categóricos. Para este fin se optó por la alternativa de reformular la búsqueda como un problema de selección de variables y para ello se usó LASSO Binomial. Usando secuencias simuladas para probar ambas metodologías se obtuvieron resultados, en términos de precisión, comparables con los de estudios existentes.

Palabras clave: Intrones y exones, proporción de nucleótidos, medidas de divergencia, selección de variables.

TWO METHODS FOR THE PREDICTION OF INTRONS AND EXONS INTO A GENE

Evelio Hernández Juárez, Dr.

Colegio de postgraduados, 2015

ABSTRACT

For eukaryotic organisms, within a gen in addition to protein coding regions (exons) there are noncoding regions know as introns. There are different approaches to find the boundaries between introns and exons, one of those tries to distinguish changes in the composition (proportion) of its nucleotides and in this context the recursive segmentation based on measures of divergence can be used. To implement this methodology a threshold (a parameter that serves as a stopping criterion) needs to be define. Here it is proposed that the choice of the threshold be through by Montecarlo distribution of the maximum divergence; moreover, that such the threshold depends on the size of the sequence under study.

Under the same perspective of finding differences in the proportion of nucleotides within the sequences, the search of change points in categorical data can be used. For this purpose, the alternative of reformulating the search as a problem of selection of variables was used following the LASSO binomial. Simulated sequences were used to test both methods and results similar to those of previous studies were obtained.

Keywords: introns and exons, nucleotide proportion, divergence measures, variable selection.

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico brindado en la realización de mis estudios.

Al Colegio de Postgraduados, por la oportunidad de continuar mi preparación personal y profesional.

Al Dr. Gustavo Ramírez Valverde, por su gran apoyo y por el tiempo dedicado.

A los integrantes del consejo particular por su orientación y apoyo en este trabajo.

A Rosario y Ernesto, todo lo que hago es por ustedes.

A mis padres y hermanos, por el apoyo y confianza que me han brindado siempre.

A mis maestros y a todas aquellas personas que de forma directa o indirecta contribuyeron a la culminación de este trabajo.

CONTENIDO

I. INTRODUCCIÓN 1

1.1. Importancia y justificación del estudio 1

1.2 Objetivos. 2

II. ADN 3

2.1. Transcripción y traducción 5

III. MÉTODOS DE SEGMENTACIÓN DE GENES 9

3.1. Métodos de segmentación basados en divergencias 9

3.2. Medidas de divergencia 11

3.3. Determinación del umbral 12

IV. PREDICCIÓN DE GENES 14

4.1. Métodos de segmentación usados para la predicción de genes 15

4.1.1. Estudio de simulación 16

4.2. Predicción de genes como un problema de búsqueda de puntos de cambio 19

4.2.1 LASSO 20

4.2.2. LASSO binomial 21

4.2.3. Estudio de simulación 24

V. RESULTADOS Y DISCUSIÓN 26

5.1. Medida de comparación 26

5.2. Método basado en divergencias 27

5.3. Método basado en la búsqueda de puntos de cambio usando LASSO binomial 29

VI. CONCLUSIONES 38

REFERENCIAS 39

LISTA DE FIGURAS

1.	El código genético ……..……...	1
2.	Transcripción de dos genes ……..……..……..	5
3.	Adición de un aminoácido a la cadena polipeptídica creciente durante la traducción del ARNm ……..……...…….……	6
4. a)	Secuencia especifica de ARNt de la alanina de levadura ……..	7
4. b)	Esquema de la estructura tridimensional real del ARNt de la fenilalanina de levadura ……..……..	7
5. a)	Transcripción y traducción en procariontes ……..……	8
5. b)	Transcripción y traducción en eucariontes ……..……..	8
6.	Densidades empíricas de la divergencia de Jensen-Shannon, para diferentes longitudes de secuencias ……..……...	18
7.	Número de variables incluidas para cada valor de λ, para secuencia 2 con un solo punto de cambio ……..……	30

LISTA DE CUADROS

1. Umbrales para diferentes longitudes de secuencias ……... 18

2. Algunos alfabetos estadísticos para secuencias de ADN ..……..…….. 24

3. Promedio de la precisión $(1 - D)$ de cada método ……...…….…… 27

4. Precisión $(1 - D)$ de cada modelo con diferente número de variables incluidas. Secuencias con un solo punto de cambio …… 31

5. Precisión $(1 - D)$ de cada modelo con diferente número de variables incluidas. Secuencias con cuatro puntos de cambio …… 33

6. Precisión $(1 - D)$ de cada modelo con diferente número de variables incluidas. Secuencias con nueve puntos de cambio …… 35

I. INTRODUCCIÓN

1.1. Importancia y justificación del estudio

El propósito de la predicción de genes es la identificación algorítmica de trozos de secuencias (usualmente de ADN), que son biológicamente funcionales. Se trata de encontrar regularidades estadísticas o patrones estructurales de organización y de función en los genomas, independientemente de que estos sean de vegetales o animales, procariotas o eucariotas.

Para organismos eucariontes, en un gen además de las regiones codificantes de proteínas (exones) se encuentran regiones no codificantes llamadas intrones. Existen diferentes enfoques para encontrar las fronteras entre intrones y exones, uno de ellos trata de distinguir cambios en la composición (proporción) de sus nucleótidos y en este contexto existen metodologías tales como la segmentación basada en medidas de divergencia y la búsqueda de puntos de cambio. Para ambas metodologías existen dificultades y variantes en sus respectivas implementaciones.

Una dificultad que conllevan los métodos basados en medidas de divergencia (como Jensen-Shannon y Jensen-Rényi) es la determinación de un umbral que sirve como criterio de parada (Bernaola-Galván, 2000), en este estudio se propone una alternativa basada en la distribución Monte-Carlo de la máxima divergencia, además de que dicho umbral dependa del tamaño de la secuencia considerada.

En lo referente a la búsqueda de puntos de cambio se optó por la alternativa que ofrecen Harchaoui y Lévy-Leduc (2008) en el sentido de reformular la búsqueda como un problema de selección de variables. Para ello se usa LASSO (Least Absolute Shrinkage eStimatOr), método de selección de variables presentado por Tibshirani en 1996, el cual minimiza la suma residual de los cuadrados, sujeto a que la suma del valor absoluto de los coeficientes sea menor que una constante t.

1.2 Objetivos.

- Comparar la eficiencia de algoritmos basados en medidas de divergencia de Jensen-Shannon y Jensen-Rényi, para la detección de cambios composicionales usando secuencias simuladas.

- Determinar si la inclusión de un umbral que depende del tamaño de la secuencia, proporciona ganancia en términos de eficiencia.

- Valorar la eficiencia del uso de LASSO Binomial para detectar puntos de cambio en secuencias simuladas de nucleótidos.

- Contrastar los métodos basados en divergencias con el uso de LASSO Binomial.

II. ADN

El ácido desoxirribonucleico, frecuentemente abreviado como ADN (y también DNA, del inglés deoxyribonucleic acid), es una macromolécula que contiene la información genética usada en el desarrollo y el funcionamiento de los organismos vivos conocidos y de algunos virus. El ADN de todos organismos esta formado por los mismos componentes llamados nucleótidos, y cada uno de estos, a su vez, está formado por un azúcar, una base nitrogenada y un grupo fosfato que actúa como enganche entre los nucleótidos. (Lewin, 2006)

Las cuatro bases nitrogenadas que se encuentran en el ADN son adenina (A), citosina (C), guanina (G) y timina (T). Ringo (2007) menciona que éstas se clasifican en dos grupos: las bases púricas o purinas (adenina y guanina) formadas por dos anillos unidos entre sí, y las bases pirimidínicas o pirimidinas (citosina y timina), derivadas de la pirimidina las cuales cuentan con un solo anillo. En los ácidos nucléicos existe una quinta base pirimidínica, denominada uracilo (U), que normalmente ocupa el lugar de la timina en el ácido ribonucleico (ARN).

La doble hélice de ADN se mantiene estable mediante la formación de puentes de hidrógeno entre las bases asociadas a cada una de las dos cadenas entrelazadas (hebras). Los puentes de hidrógeno pueden romperse y formarse de nuevo de forma relativamente sencilla; por esta razón, las dos hebras de la doble hélice pueden separarse bien por fuerza mecánica o por alta temperatura (Clausen-Schaumann, 2000).

Cada tipo de base en una hebra forma un enlace únicamente con un tipo de base en la otra hebra, lo que se denomina "complementariedad de las bases". Las purinas forman enlaces con las

pirimidinas, de forma que A se enlaza sólo con T, y C sólo con G. La organización de dos nucleótidos apareados a lo largo de la doble hélice se denomina apareamiento de bases.

Según Lewin (2006) el gen es la unidad fundamental de la herencia y cada uno de estos es parte de una secuencia continua de ADN. La secuencia de nucleótidos en el ADN es importante no por propia estructura, si no por que codifica una secuencia de aminoácidos que constituyen la correspondiente proteína. La relación entre una secuencia de ADN y la secuencia correspondiente de proteínas es llamada código genético**.**

El código genético contiene sesenta y cuatro codones, cada uno de los cuales está compuesto por tres nucleótidos contiguos en una cadena de ADN (Figura 1), sesenta y uno de estos codones especifican aminoácidos, y cada triplete codifica un único aminoácido.

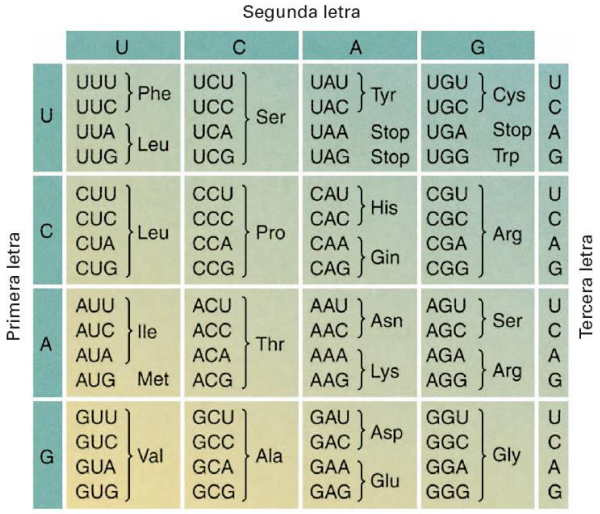

Figura 1. El código genético. Tomada de Griffiths, et al (2002).

Uno de estos tripletes (AUG) además de codificar al aminoácido metionina, señala el inicio de las secuencias de ADN que codifican proteínas. Cualquiera de los tres tripletes restantes (UAA, UAG o UGA) puede señalar el final de la cadena codificada. Algunos aminoácidos pueden ser especificados por más de un codón, pero ningún codón especifica a más de un aminoácido.

2.1. Transcripción y traducción

De acuerdo con Lewin (2006) la expresión de un gen en forma de proteína requiere que el ADN sea transcrito a ARN (Figura 2). Este proceso es catalizado por ARN polimerasas, que son enzimas que sintetizan cadenas de ARN, copiando la secuencia de nucleótidos de una de la hebras del ADN, siguiendo las reglas de apareamiento entre bases complementarias. Los genes que codifican proteínas producen ARN mensajeros (ARNm) llamados así porque transportan el mensaje contenido en el ADN e intervienen directamente en la síntesis de proteínas.

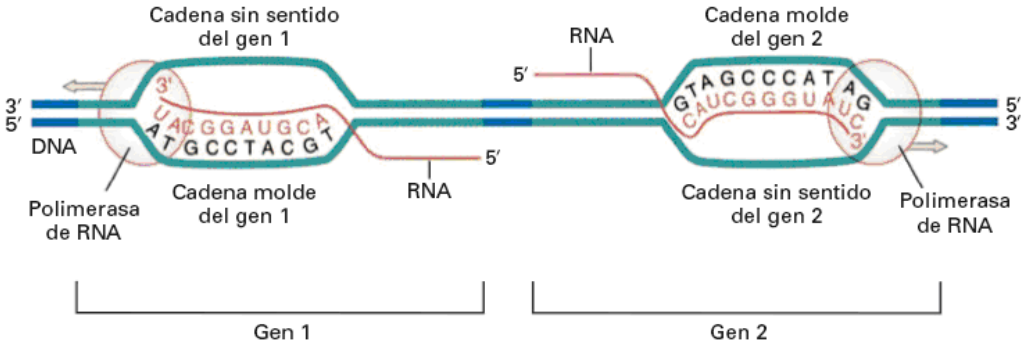

Figura 2. Transcripción de dos genes. Tomada de Griffiths, *et al* (2002).

Singer (1993) menciona que algunos genes no codifican proteínas, en lugar de ARNm, su transcripción produce moléculas de ARN que son necesarias para la maduración de los diferentes tipos de ARN y para la traducción de las secuencias de los ARNm para generar proteínas.

El proceso mediante el cual una secuencia de nucleótidos es traducida a una cadena proteica es complejo e implica un gran número de pasos iterativos.

Los ribosomas que contienen más de cincuenta proteínas diferentes y tres clases de ARN catalizan la traducción a proteínas de los ARNm. El ensamblaje de una cadena de proteínas empieza cuando los ribosomas se unen al ARNm. La cadena de proteínas se alarga aminoácido por aminoácido, uno por vez a medida que el ribosoma se desplaza codón por codón a lo largo del ARNm (Figura 3).

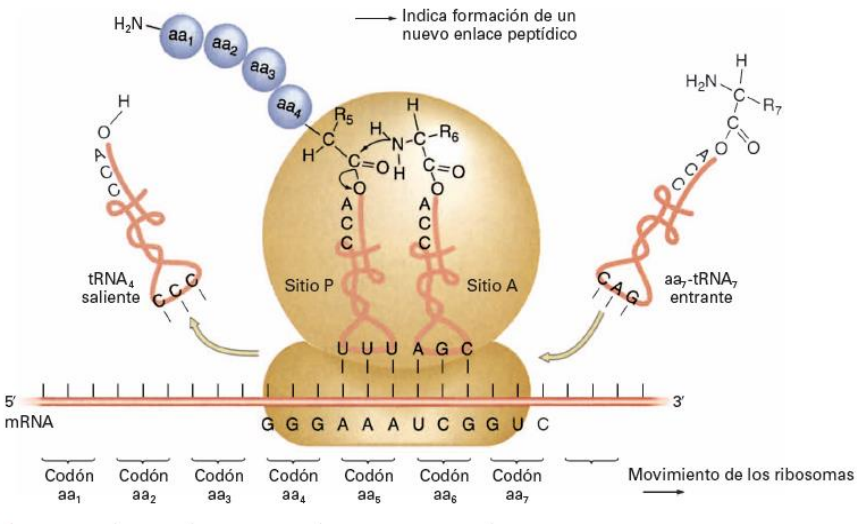

Figura 3. Adición de un aminoácido a la cadena polipeptídica creciente durante la traducción del ARNm. Tomada de Griffiths, *et al* (2002).

Cada aminoácido está unido a un ARN de transferencia (ARNt) apropiado, que contiene un triplete (anticodón) que es complementario al triplete codificante en el ARNm (Figura 4). El apareamiento entre el codón y el anticodón coloca al aminoácido correcto en su sitio y permite la unión de los nuevos aminoácidos al extremo en crecimiento de la cadena proteica. Cada vez que un ribosoma recorre la longitud total de la secuencia codificante del ARNm se sintetiza una molécula completa de la proteína correspondiente.

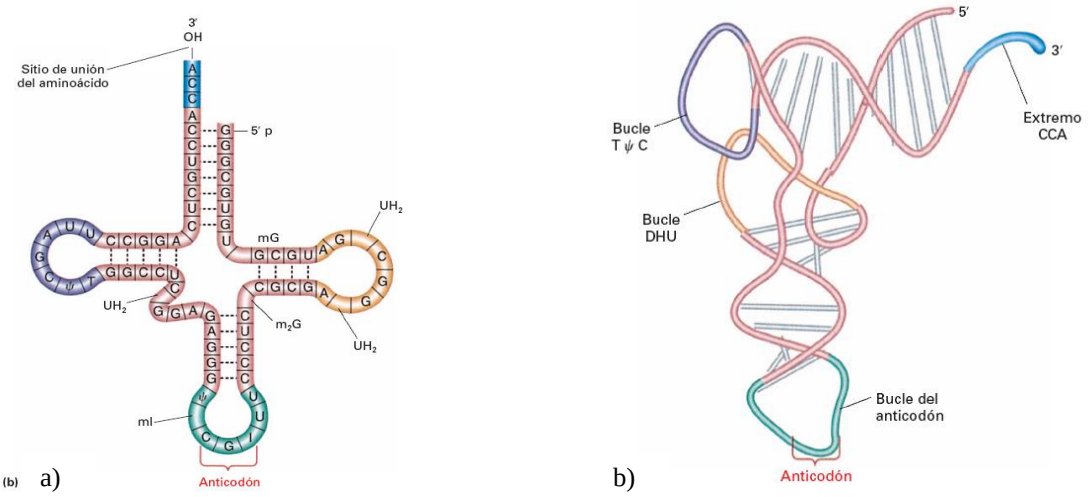

Figura 4. a) Secuencia especifica de ARNt de la alanina de levadura. b) Esquema de la estructura tridimensional real del ARNt de la fenilalanina de levadura. Tomadas de Griffiths, *et al* (2002).

Lewin (2006) menciona que en procariontes ambos procesos transcripción y traducción ocurren en el mismo sitio, ya que estos organismos carecen de un núcleo definido (Figura 5 a). En organismos eucariontes el resultado inmediato de la transcripción es llamado pre-ARNm y debe ser transportado al citoplasma para que se lleve a cabo la traducción. El pre-ARNm requiere un proceso para generar el ARNm maduro (Figura 5 b).

Un paso crucial en el proceso de la expresión de un gen es la edición del ARNm, ya que el pre-ARNm contiene regiones que no codifican proteínas (intrones). En el proceso de edición estas regiones son eliminadas y así se genera el ARNm maduro, el cual consta sólo de secuencias que codifican proteínas (exones).

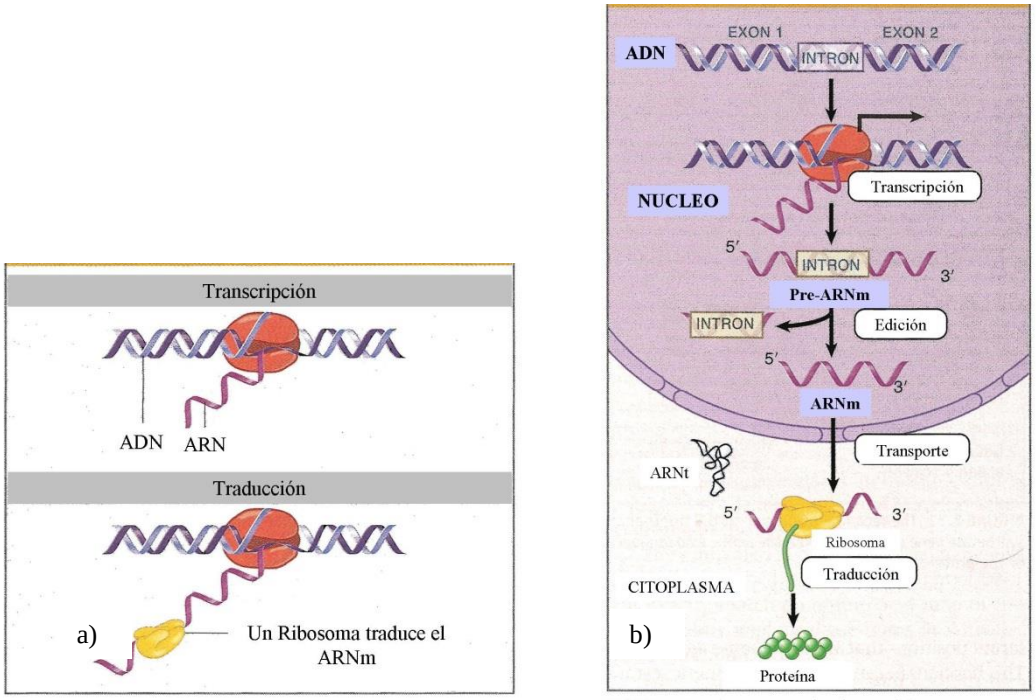

Fig. 5. a) Transcripción y traducción en procariontes. b) Transcripción y traducción en eucariontes. Tomadas de Lewin (2006).

III. MÉTODOS DE SEGMENTACIÓN DE GENES

En 1976, Bernardi y sus colaboradores propusieron una teoría sobre la composición de los genomas de vertebrados. Esta descripción, más tarde fue bautizada como la teoría "isochore". (Elhaik *et al.*, 2010). Isochores se definen como fragmentos genómicos de más de 300 kb que son relativamente homogéneos en su contenido de guanina y citosina (G-C).

Además de la ausencia de un acuerdo sobre la definición de "relativa homogeneidad", gran parte de la controversia sobre la existencia de isochores parece ser el resultado de las dificultades en la identificación de dominios de composición homogénea dentro de las secuencias del genoma.

Se han propuesto muchos métodos para la detección de isochores en secuencias genómicas. Estos métodos particionan la secuencia en dominios composicionalmente homogéneos en función de criterios predefinidos. A éstos se les conoce como métodos de segmentación. Algunos de estos métodos se resumen y comparan en Braun y Müller (1998) y en Elhaik *et al.* (2010).

3.1. Métodos de segmentación basados en divergencias

Elhaik *et al.* (2010) realizaron una comparación de 7 algoritmos de segmentación y sus resultados muestran que los algoritmos recursivos de segmentación entrópica basados en la divergencia de Jensen-Shannon superan a los otros algoritmos usados en el estudio; sin embargo,

incluso estos algoritmos funcionan mal en ciertos casos debido a la elección arbitraria de un criterio de parada.

El proceso de segmentación entrópica, particiona una secuencia heterogénea de ADN en subsecuencias homogéneas, usando una "función de contraste"; es decir, una función comparativa que alcanza valores bajos cuando se comparan dos regiones de ADN con características similares y toma valores altos si comparan dos regiones con características diferentes. Como función de contraste se suelen usar medidas de entropía, por lo que se les conoce como segmentación entrópica.

La divergencia de Jensen-Shannon (Lin, 1991) es uno de los métodos más utilizados; sin embargo, existen otras medidas de divergencia como la Jensen-Rényi, la cual, según lo reportado por Nicorici y Astola (2005), mejora la precisión de la segmentación del ADN.

Los algoritmos recursivos de segmentación barren la secuencia de ADN y calculan una medida de divergencia para cada posición i que divide la secuencia en dos subsecuencias. La posición i^* en la que la divergencia alcanza su máximo se acepta como punto de corte. Se realiza este mismo procedimiento para cada subsecuencia, y el proceso de segmentación termina cuando el valor máximo es menor que un umbral (D_C) predeterminado.

3.2. Medidas de divergencia

La divergencia Jensen-Shannon (D_{JS}) cuantifica la diferencia entre dos o más distribuciones de probabilidad, es calculada por la ecuación:

$$D_{JS} = max_i \left[H - \frac{i}{N} H_L - \frac{N-i}{N} H_R \right]$$

donde:

H es la entropía de Shannon de la secuencia completa;

H_L es la entropía de Shannon de la subsecuencia a la izquierda del punto de partición *i*;

H_R es la entropía de Shannon de la subsecuencia a la derecha del punto de partición *i*;

N es el tamaño de la secuencia completa.

La función H (entropía de Shannon), para una distribución de probabilidades $\mathbf{p} = (p_1, p_2, \ldots, p_k)$ se define como:

$$H(\mathbf{p}) = -\sum_{i=1}^{k} p_i * log_2 p_i$$

La divergencia de Jensen-Rényi (D_{JR_α}) como la divergencia Jensen-Shannon se define como una medida de similitud entre dos o más distribuciones de probabilidad, fue introducida por He *et al* (2003), y se calcula mediante la ecuación:

$$D_{JR_\alpha} = max_i \left[R_\alpha - \frac{i}{N} R_{\alpha,L} - \frac{N-i}{N} R_{\alpha,R} \right]$$

donde:

R_α es la entropía de Rényi de la secuencia completa;

$R_{\alpha,L}$ es la entropía de Rényi de la subsecuencia a la izquierda del punto de partición *i*;

$R_{\alpha,R}$ es la entropía de Rényi de la subsecuencia a la derecha del punto de partición *i*;

N es el tamaño de la secuencia completa.

La función R_α (entropía de Rényi), para una distribución de probabilidades $\mathbf{p} = (p_1, p_2, \ldots, p_k)$ y un valor fijo α se define como:

$$R_\alpha(\mathbf{p}) = \frac{1}{1-\alpha} log_2 \sum_{i=1}^{k} {p_i}^\alpha$$

para $\alpha \in (0,1)$ la entropía de Rényi es cóncava y tiende a la entropía de Shannon $H(\mathbf{p})$ cuando $\alpha \to 1$.

3.3. Determinación del umbral

Dada la naturaleza de los algoritmos recursivos de segmentación, la elección del umbral es crucial para inclusión de nuevos puntos de corte, además de ser el único criterio de parada para este tipo de algoritmos.

La manera de determinar el umbral D_C cambia dependiendo del autor, Bernaola-Galván *et al* (2000) y Cohen *et al* (2005) usaron simulación para estimar la distribución acumulativa de D_{JS} y usaron el valor correspondiente al 5% de dicha distribución. Elhaik *et al*. (2010) reporta que por comunicación personal con Tal Dagan (coautor de Cohen *et al*, 2005) utilizó como umbral el valor 5.8×10^{-5} .

Bernaola-Galván *et al*., (2000) usó un enfoque basado en una prueba de hipótesis en el cual toman la decisión en función de la probabilidad de que un valor observado de la medida de divergencia en cuestión sea un valor grande cuando la hipótesis nula (de que la secuencia es homogénea) es verdadera.

Se ha usado además el criterio "BIC", llamado así por sus iniciales en inglés "Bayesian Information Criterion" (Li, 2001 a,b y Nicorici y Astola, 2005). El criterio BIC es usado en un marco de selección de modelos, y la decisión se hace ponderando el ajuste con el número de parámetros que contiene, prefiriendo modelos con menor valor de BIC, el cual se define mediante la siguiente ecuación:

$$BIC = -2 \cdot \ln\left(\hat{L}\right) + k \cdot \ln(n)$$

Donde

k es el numero de variables incluidas en el modelo M.

n es el numero de observaciones

$\hat{L}$ es el máximo valor de la función de verosimilitud del modelo M, es decir, $\hat{L} = p\left(x\middle|\hat{\theta},\ M\right)$.

$\hat{\theta}$ son los valores de los parámetros que maximizan la función de verosimilitud.

x son los datos observados.

IV. PREDICCIÓN DE GENES

El reconocimiento computacional de genes es uno de los retos en el análisis de nuevos genomas secuenciados, lo cual es fundamental para la genómica moderna, cuyo objetivo es la búsqueda de los diferentes elementos funcionales que constituyen las secuencias de ADN.

Hay dos problemas básicos en la búsqueda de genes: 1) la detección de los sitios funcionales de los genes, y 2) la detección de las regiones que codifican para las proteínas. El enfoque de este estudio es sobre el problema 2).

Existen programas computacionales para la predicción de genes, algunos de ellos disponibles en internet GENSCAN (Burge, 1997), HMMgene (Krogh, 1997), GeneMark.hmm (Lukashin y Borodovski, 1998), AUGUSTUS (Stanke y Waack 2003), SGP2 (Parra *et al*. 2003), la mayoría basados en Modelos de Markov Ocultos (HMM por sus siglas en inglés) y requieren de bases de datos con información previa de genomas similares para entrenamiento de los algoritmos; además usan una variedad de información biológica como posibles señales de secuencias implicadas en la especificación de un gen o búsquedas en bases de datos de similitud de secuencias, estas señales se utilizan principalmente para obtener fronteras probables de codificación de las regiones, y deben obtenerse a partir de la información biológica elaborada, que es altamente dependiente del genoma particular considerado.

4.1. Métodos de segmentación usados para la predicción de genes

Si bien, existen muchos métodos para detección de regiones codificantes de proteínas (exones) y regiones no codificantes (intrones), una limitación importante de la mayoría de ellos es la necesidad de "formación previa" ; es decir se necesitan grandes conjuntos de secuencias con características similares para el "entrenamiento" de dichos métodos.

Por lo tanto, se necesitan estrategias computacionales para encontrar genes que no requieran formación previa en conjuntos de datos específicos del organismo, y el proceso de segmentación entrópica puede servir como un primer paso en esta dirección. (Bernaola-Galván *et al* 2000).

Estudios como los de Bernaola-Galván *et a.l* (2000) y Nicorici y Astola (2005) reportan buenos resultados al usar segmentación entrópica en la diferenciación de intrones y exones.

El objetivo de la primera parte de este estudio es comparar la eficiencia (entendida como la detección de cambios en la composición de nucleótidos dentro de secuencias genómicas) de algoritmos basados en medidas de divergencia, Jensen-Shannon y Jensen-Rényi, así como la determinación del umbral.

Para la determinación del umbral se utiliza el enfoque de prueba de hipótesis, para lo cual es necesario conocer la distribución de las medidas de divergencia observadas bajo la hipótesis nula (secuencias sin cambios composicionales), sin embargo, la forma exacta de esta distribución no es fácil de obtener (Pettitt, 1980). Una aproximación asintótica es conocida (Horváth, 1989), pero para el caso de identificación de regiones codificantes se tienen secuencias de menor tamaño

por lo que la aproximación asintótica es cuestionable. Bernaola-Galván *et al*. (2000) proponen una aproximación basada en la distribución Monte Carlo del valor del máximo de la divergencia con un solo punto de corte.

En este estudio se propone un umbral dinámico, que varíe dependiendo de la longitud de la secuencia; los valores del umbral se estiman con la distribución Monte Carlo del valor del máximo de cada medida de divergencia respectiva (Jensen-Shannon y Jensen-Rényi).

4.1.1. Estudio de simulación

Con la finalidad de centrar el estudio en la capacidad de los algoritmos para identificar cambios en la composición (proporción) de nucleótidos, se han simulado secuencias genómicas que contienen un número predeterminado de segmentos con composición homogénea de nucleótidos. Con este procedimiento no se pretende simular toda la complejidad subyacente en las secuencias genómicas, sin embargo algoritmos que no detecten los cambios en un escenario como este no podrán competir en secuencias genómicas reales.

Todas las secuencias simuladas fueron de tamaño 2000, tuvieron diferente número de subsecuencias de composición homogénea: 2, 5 y 10; las subsecuencias se construyeron con longitudes iguales.

Las secuencias con un solo punto de corte; es decir las que cuentan con sólo dos subsecuencias tienen como objetivo conocer si los algoritmos detectan falsos positivos, mientras

que las secuencias con mayor número de cambios de composición tienen como objetivo evaluar si los algoritmos detectan los cambios en subsecuencias cortas. Cada situación simulada se repitió diez veces.

Para cada secuencia se usó un algoritmo de segmentación recursiva, basado en la divergencia de Jensen-Shannon y otro basado en la divergencia de Jensen-Rényi, para ambos casos se evaluó un umbral único y un umbral dinámico que cambia de acuerdo con el tamaño de la subsecuencia.

Para construir la distribución Monte Carlo del máximo de la discrepancia, bajo la hipótesis nula, se generaron 10,000 secuencias de tamaño 2000 (ya que este es el tamaño de todas las secuencias simuladas usadas en este estudio), cada una de ellas con diferente función de distribución la cual fue generada de forma aleatoria, y se calculó el valor de D_{JS}, un procedimiento análogo se realizó para D_{JR_α}, para un $\alpha = 0.8$. El umbral se puede definir como el percentil 0.95 de la distribución Monte Carlo. El mismo procedimiento se realizó para secuencias de tamaño 1500, 1000, 500, 200 y 100.

En la Figura 6 se muestran tres diferentes densidades empíricas de la máxima divergencia de Jensen-Shannon.

En el Cuadro 1 se presentan los valores de los umbrales obtenidos para cada longitud y cada medida de divergencia. Como puede observarse, existen marcadas diferencias, sobre todo para longitudes pequeñas, por lo que utilizar un solo valor para el umbral puede llevar a resultados muy distintos.

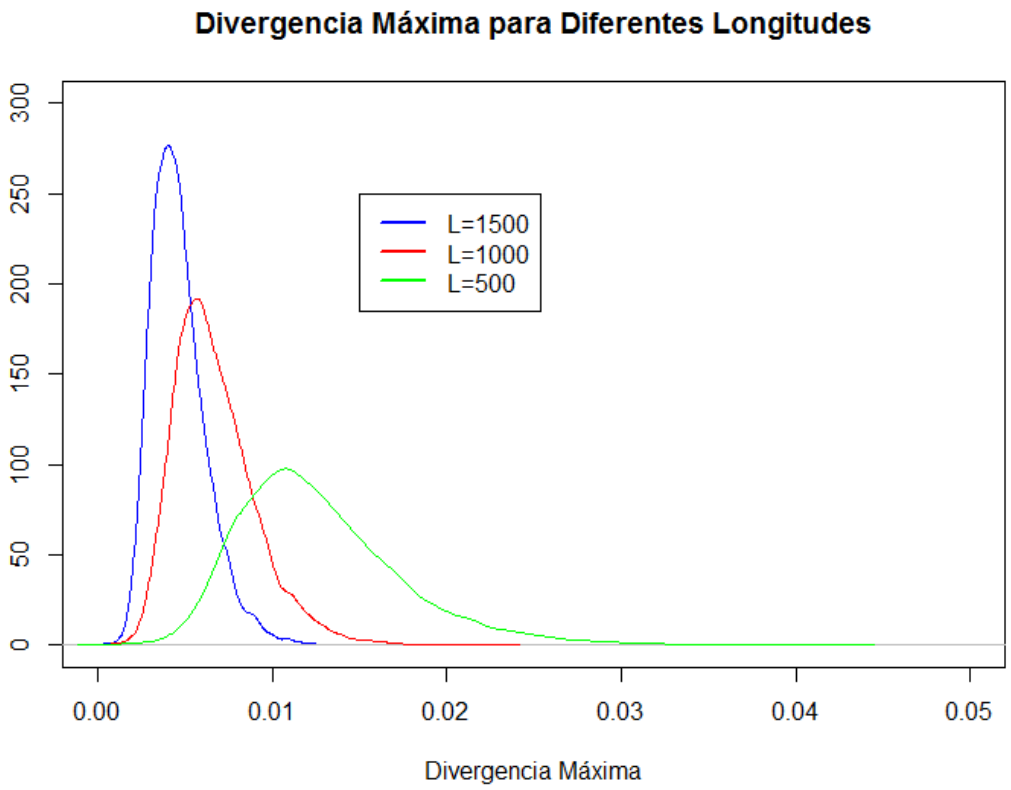

Figura 6. Densidades empíricas de la divergencia de Jensen-Shannon, para diferentes longitudes de secuencias.

Cuadro 1. Umbrales para diferentes longitudes de secuencias.

Longitud	**Umbral DJS**	**Umbral DJR**
2000	0.0057	0.0061
1500	0.0075	0.0081
1000	0.0111	0.0121
500	0.0215	0.0230
200	0.0499	0.0558
100	0.0914	0.1065

4.2. Predicción de genes como un problema de búsqueda de puntos de cambio

En Braun y Müller (1998) se muestra que la mayoría de los enfoques de segmentación se pueden incluir en una versión adecuada del problema de múltiples puntos de cambio. Si se considera una secuencia de ADN como una sucesión de variables $Y_1, Y_2 \ldots, Y_n$ donde Y_i toma uno de los cuatro valores A, C, G o T; se supone que hay segmentos dentro de los cuales las observaciones siguen la misma o casi la misma distribución (proporción de nucleótidos) y entre los cuales las observaciones tienen diferentes distribuciones. Así, las observaciones $Y_1, Y_2 \ldots, Y_n$ se dividen en $R + 1$ segmentos contiguos y entonces los puntos de cambio corresponden con los puntos finales dichos segmentos.

Para la búsqueda de puntos de cambio se usó el enfoque presentado por Harchaoui y Levy-leduc (2008), el cual consiste en la reformulación de esta tarea en un contexto de selección de variables ficticias que representan todas las posibles ubicaciones de los puntos de cambio.

Si se considera el modelo:

$$Y = X\beta + \varepsilon$$

Donde Y es un vector de observaciones de dimensión $n \times 1$, X es una matriz triangular inferior de dimensión $n \times n$ con los elementos no-ceros iguales a 1 y ε es un vector cuyas entradas son variables aleatorias independientes e idénticamente distribuidas con media cero y varianza finita. El vector β tendrá todos sus componentes iguales a cero, excepto los correspondientes a los instantes de los puntos de cambio.

Harchaoui y Levy-leduc (2008) demostraron que el problema de la estimación de puntos de cambio es equivalente a

$$\min_{\beta} \ \|\mathrm{Y} - \mathrm{X}\beta\|_n^2 \quad sujeto\ a \quad \|\beta\|_1 \leq s$$

Donde $\|u\|_1$ y $\|u\|_n$ se definen para un vector $u = (u_1, u_2, \dots, u_n) \in \mathbb{R}^n$ como

$\|u\|_1 = \sum_{j=1}^{n} |u_j|$ y $\|u\|_n^2 = \frac{1}{n}\sum_{j=1}^{n} u_j^2$ respectivamente.

Lo anterior es equivalente al problema resuelto por LASSO. El algoritmo LAR, como se describe en Efron *et al.* (2004) proporciona un algoritmo eficiente para calcular la ruta completa de regularización para el problema LASSO.

4.2.1 LASSO

LASSO (Least Absolute Shrinkage and Selection Operator), es un método de regresión restringida, el cual minimiza la suma residual de cuadrados sujetos a que la suma del valor absoluto de los coeficientes sea menor que una constante. (Tibshirani, 1996); es decir, LASSO resuelve:

$$\min_{\beta} \sum_{i}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Sujeto a

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

O equivalentemente, minimizando:

$$\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{y_i}-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

Donde t, $\lambda > 0$ son los respectivos parámetros de penalización. Para valores crecientes de λ o decrecientes de t, los coeficientes β_j se contraen hacia cero.

LASSO fue en un inicio concebido para variables continuas, sin embargo en Friedman *et al*. (2010) se presenta una aproximación para el modelo lineal general y además un paquete para el software R llamado "glmnet".

4.2.2. LASSO binomial

Cuando la variable respuesta es binaria, a menudo se usa el modelo de regresión logística. Si la variable respuesta se denota por "G", tomando valores en $G = 0,1$ (el etiquetado de los elementos es arbitrario). El modelo de regresión logística representa las probabilidades condicionales a través de una función lineal de los predictores

- $P(G = 1|x) = \frac{1}{1+e^{-(\beta_0+x'\beta)}}$

- $P(G = 0|x) = 1 - P(G = 1|x) = 1 - \frac{1}{1+e^{-(\beta_0+x'\beta)}} = \frac{e^{-(\beta_0+x'\beta)}}{1+e^{-(\beta_0+x'\beta)}} = \frac{1}{1+e^{(\beta_0+x'\beta)}}$

Esto implica que:

$$\ln\left[\frac{P(G=1|x)}{P(G=0|x)}\right] = \beta_0 + x'\beta$$

Para este caso se ajusta el modelo mediante máxima verosimilitud regularizada. Sea $p(x_i) = P(G=1|x_i)$ la probabilidad para la *i-esima* observación en un valor particular para los parámetros $(\beta_0\,, \beta)$, entonces se maximiza la log-verosimilitud penalizada:

$$\max_{(\beta_0\,,\beta)\in\mathbb{R}^{p+1}} \left[\frac{1}{N}\sum_{i=1}^{N}\{I(g_i=1)\cdot \ln p(x_i) \;+\; I(g_i=0)\cdot \ln\big(1-p(x_i)\big)\} - \lambda P_\alpha(\beta)\right]$$

Denotando $y_i = I(g_i = 1)$, la log-verosimilitud en la expresión anterior se puede escribir en forma más explicita como

$$l(\beta_0\,, \beta) = \frac{1}{N}\sum_{i=1}^{N}\{y_i \cdot (\beta_0 + x'\beta) - ln\big(1 + e^{(\beta_0 + x'\beta)}\big)\}$$

la cual es una función cóncava de los parámetros. El algoritmo de Newton para maximizar la log-verosimilitud (no penalizada) es equivalente a mínimos cuadrados ponderados iterativos.

Por tanto, si las estimaciones actuales de los parámetros son $(\tilde{\beta}_0, \tilde{\beta})$, se forma una aproximación cuadrática para la log-verosimilitud (expansión de Taylor sobre las estimaciones actuales), que es:

$$l_Q(\beta_0\,, \beta) = -\frac{1}{2N}\sum_{i=1}^{N} w_i(z_i - \beta_0 + x_i{}'\beta)^2 + C\big(\tilde{\beta}_0, \tilde{\beta}\big)^2$$

Donde

$$z_i = \tilde{\beta}_0 + x_i{}'\tilde{\beta} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)\big(1-\tilde{p}(x_i)\big)}$$

$$w_i = \tilde{p}(x_i)\big(1 - \tilde{p}(x_i)\big)$$

y $\tilde{p}(x_i)$ se evalúa en los parámetros actuales. El último término es una constante. La actualización de Newton se obtiene minimizando l_Q.

La aproximación que se presenta en Friedman, *et al*. (2010), es similar. Para cada valor λ (parámetro de penalización) se crea un bucle (loop) exterior que calcula la aproximación cuadrática l_Q sobre los parámetros actuales $\big(\tilde{\beta}_0, \tilde{\beta}\big)$. Luego se usa "coordinate descent" para resolver el problema de mínimos cuadrados ponderados restringidos.

$$\min_{(\beta_0\,,\beta)\in\mathbb{R}^{p+1}} \big[-l_Q(\beta_0\,,\beta) + \lambda P_\alpha(\beta)\big]$$

Esto equivale a una secuencia de bucles anidados:

Bucle externo: Disminuir λ.

Bucle medio: Actualizar la aproximación cuadrática l_Q usando los parámetros actuales $\big(\tilde{\beta}_0, \tilde{\beta}\big)$.

Bucle interior: Usar el algoritmo "coordinate descent" para obtener un resultado de la minimización.

Con este procedimiento se logra una actualización de los parámetros, es decir los nuevos $\big(\tilde{\beta}_0, \tilde{\beta}\big)$.

En Friedman, et al. (2010) se menciona que este algoritmo no implementa ningún control de divergencia y cuando se utiliza como se recomienda no se considera necesario, por lo general hace a las aproximaciones cuadráticas muy precisas y no se ha encontrado ningún problema de divergencia hasta ahora.

4.2.3. Estudio de simulación

Con el propósito de que los dos métodos puedan ser comparados, para esta parte del estudio se usaron las mismas secuencias simuladas descritas en el apartado 4.1.1. Dichas secuencias son multinomiales ya que simulan secuencias de ADN; es decir, se representan como una secuencia $Y_1, \ldots, Y_n$, donde Y_i toma uno de los cuatro valores del alfabeto del ADN (A, C, G o T).

Los nucleótidos pueden ser clasificados en varios grupos en función de sus propiedades físicas y químicas. El cuadro 2, tomado de Braun and Müller (1998) ofrece algunos ejemplos de estas clasificaciones.

Cuadro 2. Algunos alfabetos estadísticos para secuencias de ADN.

Apareamiento	**Alfabeto**
Purina *vs*. Pirimidina	R (A o G); Y(C o T)
Fuerte *vs*. Débil	S (C o G); W (A o T)
Keto *vs*. Amino	K (T o G) ; M (A o C)

Usando la clasificación Fuerte *vs* Débil se transforma cada secuencia original en una transformada a datos binarios.

Con cada una de estas secuencias transformadas se usa regresión LASSO binomial con el paquete glmnet del software R, aquellas posiciones de las variables que resulten significativas se aceptarán como puntos de cambio, lo que a su vez indicará los fronteras entre regiones con diferentes composiciones de nucleótidos.

Debe considerarse que uno de los principales retos de LASSO es la elección adecuada del parámetro λ, el paquete glmnet elige una serie valores y para cada uno de ellos se obtiene un modelo, en el cual se incluyen las variables que son significativas.

En este trabajo se incluyeron los modelos correspondientes a los primeros 20 valores de λ usados por el paquete y se consideraron sólo los modelos con diferente número de variables incluidas, ya que el hecho de tener dos modelos con diferentes valores de λ no implica necesariamente que dichos modelos incluyan distinto numero de variables.

V. RESULTADOS Y DISCUSIÓN

5.1. Medida de comparación

Para determinar la eficiencia de los algoritmos propuestos se uso una función que mide la discrepancia entre los verdaderos sitios de cambio de composición y los puntos encontrados por cada uno de los métodos, dicha función fue introducida por Bernaola-Galván *et al.* (2000):

$$D = \frac{1}{2}\left(\sum_{i}\frac{\min_j|b_i - c_j|}{N_T} + \sum_{j}\frac{\min_i|b_i - c_j|}{N_T}\right)$$

Donde $\{b_i\}$ es el conjunto de todas las fronteras (verdaderas) entre regiones codificantes y no codificantes, $\{c_j\}$ es el conjunto de todos los cortes producidos por la metodología a evaluar y NT es la longitud total de la secuencia.

La primera suma mide la discrepancia entre los cortes y las fronteras, mediante la adición para cada frontera real de la distancia hacia el corte más cercano. La segunda sumatoria realiza la misma operación, pero ahora incluyendo para cada corte la distancia a la frontera verdadera más cercana.

Ambas sumas están obligadas a incluir no sólo la corrección de la posición de los cortes (*D* será igual a cero en el momento en que los cortes y las fronteras coincidan), sino también la diferencia entre el número de fronteras y cortes.

El valor D puede ser visto como un promedio del error en la determinación de los límites correctos entre las regiones codificantes y no codificantes, por lo tanto "$1 - D$" es una medida razonable de la precisión del método. (Bernaola-Galván *et al.*, 2000).

5.2. Método basado en divergencias

Una vez obtenidos los puntos de corte estimados por cada uno de los algoritmos recursivos y usando ambas maneras de determinar los umbrales, se calculó la precisión de cada método con $1 - \mathrm{D}$. En el Cuadro 3 se concentran los promedios resultantes de las 10 repeticiones de cada alternativa; puede observarse, que los valores $1 - \mathrm{D}$ con un umbral fijo oscilan entre 0.68 y 0.75, siendo similares a los reportados por Bernaola-Galván *et al*. (2000), que encuentra valores entre 0.6 y 0.8 trabajando con secuencias completas de tres bacterias (*Rickettsia prowazekii*, *Escherichia coli*, y *Methanococcus jannaschii*).

Cuadro 3. Promedio de la precisión ($1 - D$) de cada método.

		Subsecuencias		
		Dos	**Cinco**	**Diez**
Umbral único	DJS	0.7414	0.6840	0.7382
	DJR	0.6963	0.6802	0.7546
Umbral dinámico	DJS	0.9907	0.9601	0.9439
	DJR	0.9950	0.9630	0.6707

La ganancia obtenida al usar un umbral determinado por la distribución del valor máximo de la divergencia que dependa del tamaño de la secuencia, en lugar de usar un umbral fijo para cada

paso del algoritmo es substancial para ambas medidas de divergencia, sobre todo para secuencias cortas.

Los valores de precisión obtenidos con un umbral dinámico mostraron un mejor comportamiento prácticamente en todos los casos ya que se obtuvieron valores de precisión mayores a 0.94; con excepción del caso donde se utiliza la divergencia de Jensen-Rényi con 10 subsecuencias.

El uso de un umbral fijo es ampliamente usado en la secuenciación se ADN donde los fragmentos estudiados constan de varios miles de pares de bases (kb), por lo que no representa un problema en ese contexto. Sin embargo en la detección de exones e intrones dentro de un gen, puede ser fundamental por la variabilidad de las dimensiones de ambos tipos de segmentos.

Al comparar las dos medidas de divergencia, se observó que ambas presentan comportamientos similares en la detección de puntos de corte; sin embargo, la divergencia de Jensen-Rényi es más restrictiva para secuencias de longitud corta, lo cual lleva a omitir cambios verdaderos de composición con mayor frecuencia.

En casos en los que las subsecuencias son de mayor tamaño, esta misma característica de ser restrictiva podría llevar a la divergencia de Jensen-Rényi a ser más eficiente que la divergencia de Jensen-Shannon en lo referente a la inclusión de falsos puntos de cambio (falsos positivos).

5.3. Método basado en la búsqueda de puntos de cambio usando LASSO binomial

Una vez obtenidos los diferentes modelos correspondientes a los primeros 20 valores de λ, las variables ficticias incluidas en estos modelos indican las posiciones que se aceptan como puntos de corte y se calcula la precisión usando $1 - \mathrm{D}$.

A continuación se presenta el valor de $1 - \mathrm{D}$ obtenido con cada modelo, así como el número de variables incluidas en cada uno de estos.

En el Cuadro 4 se encuentran las 10 repeticiones de las secuencias que incluyen un solo punto de cambio (dos subsecuencias).

El hecho de que en cada secuencia los modelos obtenidos con los primeros 20 valores de λ no incluyan a muchas variables indica que al usar esta metodología no incluirá excesivos falsos puntos de corte. En la Figura 7 se muestra el número de variables incluidas en cada modelo obtenido con los diferentes valores de λ, para la secuencia numero 2.

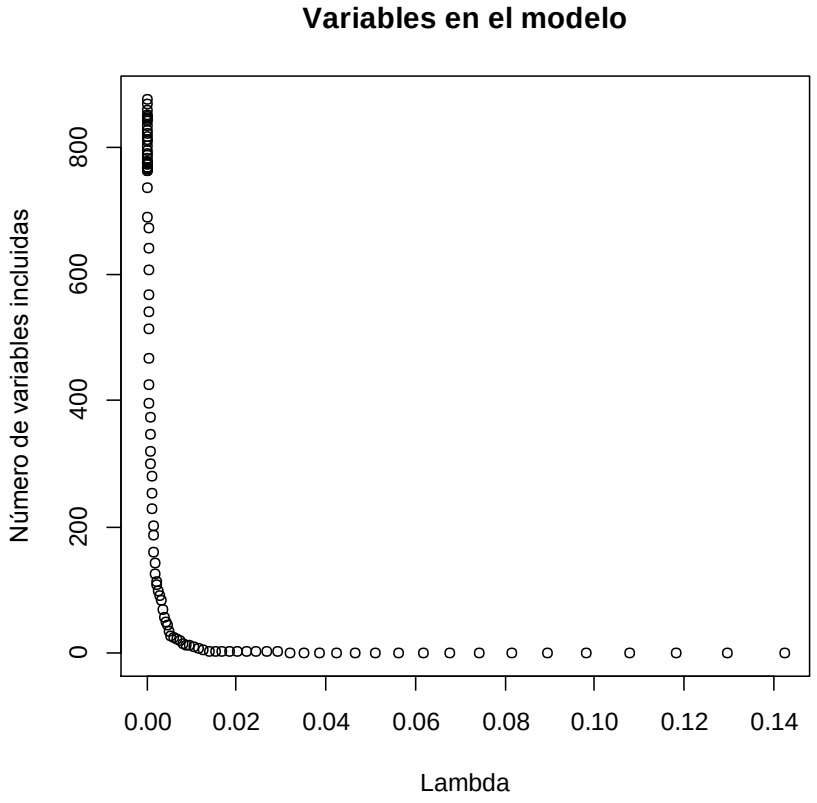

Figura 7. Número de variables incluidas para cada valor de λ, para secuencia 2 con un solo punto de cambio.

Para la secuencia 3 puede notarse que el mejor valor de precisión se obtiene con el modelo que incluye 4 variables, lo cual puede parecer inadecuado; sin embargo, esto se debe a la inclusión de una variable cercana al verdadero punto de cambio que se encuentra en la posición 1000. Las variables incluidas en el modelo con 3 son las correspondientes a las posiciones 941, 973 y 987 y para el modelo con 4 variables incluidas son la 941, 973, 987 y 1002.

Si se toma el mejor valor de $1 - D$ para cada secuencia el promedio de la precisión es 0.9924, dicho resultado es comparable con lo obtenido con los métodos basados en divergencia.

Cuadro 4. Precisión $(1 - D)$ de cada modelo con diferente número de variables incluidas. Secuencias con un solo punto de cambio.

Secuencias con un punto de cambio					
Secuencia 1		**Secuencia 2**		**Secuencia 3**	
1-D	**Variables incluidas**	**1-D**	**Variables incluidas**	**1-D**	**Variables incluidas**
0.989	1	0.997	1	0.97175	2
0.9945	2	0.92025	2	0.972	3
0.9845	3			0.97425	4
0.9695	4			0.88675	6
0.7217	5				
Secuencia 4		**Secuencia 5**		**Secuencia 6**	
1-D	**Variables incluidas**	**1-D**	**Variables incluidas**	**1-D**	**Variables incluidas**
0.9995	1	0.991	1	0.99075	2
0.99925	2	0.995	2	0.98075	3
0.868	4			0.97025	4
				0.9555	5
				0.55275	7

Cuadro 4 (continuación). Precisión $(1 - D)$ de cada modelo con diferente número de variables incluidas. Secuencias con un solo punto de cambio.

Secuencias con un punto de cambio					
Secuencia 7		**Secuencia 8**		**Secuencia 9**	
1-D	**Variables incluidas**	**1-D**	**Variables incluidas**	**1-D**	**Variables incluidas**
0.999	1	0.993	2	0.999	1
0.9952	2	0.993	3	0.7575	2
0.9587	3	0.9797	4	0.751	3
0.8827	4	0.9337	5		
0.6917	5	0.8512	6		

Secuencia 10	
1-D	**Variables incluidas**
0.9965	1
0.9917	2
0.7497	3
0.5115	4

De manera análoga al caso de las secuencias con un punto de cambio, en el Cuadro 5 se presentan los valores de $1 - D$ correspondientes a los modelos que incluyen diferente número de variables, esto para las secuencias que incluyen 4 puntos de cambio.

El promedio de los mejores valores de precisión para secuencia es 0.9234, lo cual una vez más es comparable con los valores obtenidos con las mejores opciones de los métodos basados en de divergencias.

Cuadro 5. Precisión $(1 - D)$ de cada modelo con diferente número de variables incluidas. Secuencias con cuatro puntos de cambio.

Secuencias con cuatro puntos de cambio					
Secuencia 1		**Secuencia 2**		**Secuencia 3**	
1-D	**Variables incluidas**	**1-D**	**Variables incluidas**	**1-D**	**Variables incluidas**
0.79375	2	0.3925	1	0.77975	2
0.89325	3	0.37625	2	0.7695	3
0.98475	5	0.669	3	0.8725	4
0.98575	6	0.705	4	0.77375	5
0.917	8	0.80425	7	0.8735	7
Secuencia 4		**Secuencia 5**		**Secuencia 6**	
1-D	**Variables incluidas**	**1-D**	**Variables incluidas**	**1-D**	**Variables incluidas**
0.39625	2	0.779	3	0.6637	3
0.75475	4	0.7845	5	0.698	5
0.85225	6	0.884	6	0.617	7
0.95275	7	0.97675	7	0.726	8
0.94875	8	0.9725	8	0.8222	9

Cuadro 5 (continuación). Precisión $(1 - D)$ de cada modelo con diferente número de variables incluidas. Secuencias con cuatro puntos de cambio.

Secuencias con cuatro puntos de cambio					
Secuencia 7		**Secuencia 8**		**Secuencia 9**	
1-D	**Variables incluidas**	**1-D**	**Variables incluidas**	**1-D**	**Variables incluidas**
0.7712	3	0.76175	2	0.7875	2
0.8675	6	0.761	4	0.8882	4
0.9535	7	0.797	8	0.982	7
0.935	13	0.888	11	0.9767	9
0.879	14	0.89	12	0.9797	10

Secuencia 10	
1-D	**Variables incluidas**
0.75475	3
0.7705	4
0.77	5
0.996	9
0.8857	10

En el Cuadro 6 se presentan los valores de $1 - D$ correspondientes a los modelos para las secuencias que incluyen 9 puntos de cambio. Estas secuencias con segmentos cortos son particularmente complicadas para todos los métodos, aquí el promedio de las mejores precisiones es de 0.8313. Este valor, si bien es menor que el obtenido con el método basado en la divergencia de Jensen-Shannon usando un umbral dinámico, es superior al obtenido con la divergencia de Jensen-Rényi usando también un umbral dinámico.

Cuadro 6. Precisión $(1 - D)$ de cada modelo con diferente número de variables incluidas. Secuencias con nueve puntos de cambio.

SECUENCIAS CON NUEVE PUNTOS DE CAMBIO					
Secuencia 1		**Secuencia 2**		**Secuencia 3**	
1-D	**Variables incluidas**	**1-D**	**Variables incluidas**	**1-D**	**Variables incluidas**
0.20075	2	0.1905	2	0.20275	2
0.49975	3	0.39625	4	0.46475	5
0.4905	5	0.53775	5	0.63	9
0.73	6	0.77825	7	0.72825	10
0.7785	7	0.89175	11	0.72925	12
0.8665	8	0.94025	12	0.6395	13
0.85875	12	0.892	13	0.73175	14
Secuencia 4		**Secuencia 5**		**Secuencia 6**	
1-D	**Variables incluidas**	**1-D**	**Variables incluidas**	**1-D**	**Variables incluidas**
0.178	2	0.1855	2	0.5995	3
0.58375	3	0.19975	3	0.6667	6
0.77825	5	0.37675	5	0.7647	9
0.91975	12	0.53025	6	0.7617	12
0.95075	15	0.67825	11	0.765	16
0.911	17	0.77125	12	0.8105	17
0.90525	19	0.6435	15	0.8095	20

Cuadro 6 (continuación). Precisión $(1 - D)$ de cada modelo con diferente número de variables incluidas. Secuencias con nueve puntos de cambio.

SECUENCIAS CON NUEVE PUNTOS DE CAMBIO					
Secuencia 7		**Secuencia 8**		**Secuencia 9**	
1-D	**Variables incluidas**	**1-D**	**Variables incluidas**	**1-D**	**Variables incluidas**
0.19025	3	0.1647	2	0.3662	4
0.7717	7	0.5907	6	0.735	8
0.82075	8	0.6552	13	0.7667	10
0.8485	10	0.7992	17	0.8682	12
0.7505	13	0.848	18	0.8645	16
0.8007	18	0.81975	19	0.8222	17
0.7262	21	0.7927	20	0.7542	21

Secuencia 10	
1-D	**Variables incluidas**
0.4565	4
0.35275	7
0.4207	10
0.5227	12
0.4835	13
0.6295	20
0.678	22

Es importante mencionar que para la mayoría de las secuencias existen casos en los que las variables incluidas en los modelos son ser cercanas entre si por regiones, por ejemplo, para la secuencia 2, en el modelo que incluye 12 variables, éstas son: 196, 201, 387, 406, 596, 800, 1003 1201, 1386, 1559, 1803, 1846, donde es se puede notar la cercanía entre algunas de ellas, en

particular 196 con 201 y 387 con 406. Lo anterior indica que en esa región existe evidencia de un cambio en la composición de nucleótidos.

VI. CONCLUSIONES

- Una combinación del algoritmo recursivo de segmentación basado en la divergencia de Jensen-Shannon y el uso de un umbral dinámico, que dependa del tamaño de la secuencia considerada, ofrece la mejor alternativa entre las opciones de algoritmos basados en divergencias presentadas en este estudio.

- Los puntos de cambio determinados por LASSO binomial, así como la cantidad de ellos es adecuada considerando los verdaderos puntos de cambio; es decir, en general el método no incluye falsos positivos al menos para los primeros valores del parámetro λ.

- La determinación del valor adecuado del parámetro para LASSO es la principal dificultad para la implementación de este algoritmo, sin embargo con el mejor resultado de precisión medida con la función $1 - D$ los resultados son similares a los obtenidos con los métodos basados en divergencias.

REFERENCIAS

Bernaola-Galván P., I. Grosse,P. Carpena,J.L. Oliver,R. Róman-Roldán and H.E.Stanley. 2000. Finding borders between coding and noncoding DNA regions by an entropic segmentation method. Physical Review Letters vol.85 no.6 p. 1342:1345.

Braun J. V. and H. G. Müller. 1998. Statistical methods for DNA sequence segmentation. Statistical Science 13: 142-162.

Burge C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology 268: 78:94.

Clausen-Schaumann H., M. Rief, C. Tolksdorf and H.E. Gaub. 2000. Mechanical stability of single DNA molecules. Biophys J 78 (4): 1997–2007. PMID 10733978.

Cohen N., T. Dagan, L. Stone and D. Graur. 2005. GC composition of the human genome: in search of isochores. Molecular Biology and Evolution, 22: 1260-1272.

Efron, B., T. Hastie, I. Johnstone and R. Tibshirani. 2004. Least angle regression. The Annals of Statistics, 32: 407-499.

Elhaik E., D. Graur and K. Josić (2010) Comparative testing of DNA segmentation algorithms using benchmark simulations. Molecular Biology and Evolution 27: 1015-1024.

Friedman J., T. Hastie and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33:1-22.

Griffiths, A. J. F., J.H. Miller, D.T. Suzuki, R. C. Lewontin and W. M. Gelbart. 2002. Genética. Séptima Edición. Mc Graw-Hill Interamericana. Madrid, España. 860 p.

He Y., A. B. Hamzaand H. Krim (2003) A generalized divergence measure for robust image registration. IEEE Transactions on Signal Processing 51: 1211-1220.

Horváth A. L. 1989. The limit distributions of likelihood ratio and cumulative sum tests for a change in a binomial probability. Journal of Multivariate Analysis 31: 148-159.

Krogh, A. 1997. Two methods for improving performance of an HMM and their application for gene-finding. *In:* Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology. Gaasterland, T., P. Karp, K. Karplus, C. Ouzounis, C. Sander, and A. Valencia (eds.). American Association for Artificial Intelligence. Menlo Park, California. pp:179-186.

Harchaoui, Z. and C. Levy-Leduc. 2008. Catching change-points with lasso. Advances in Neural Information Processing Systems 20: 161-168.

Lewin, B. 2006. Genes IX. Jones and Bartlett Publishers. Sudbury, Massachusetts. 892 p.

Li W. 2001a. New stopping criteria for segmenting DNA sequences. Physical Review Letters 86: 5815-5818.

Li W. 2001b. DNA segmentation as a model selection process. *In*: Proceedings of the Fifth Annual International Conference on Computational Biology. Association for Computing Machinery Press. New York. pp: 204-210.

Lin J. 1991. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory 37: 145-151.

Lomsadze A., V. Ter-Hovhannisyan, Y. O. Chernoff and M. Borodovsky. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Research 33: 6494-6506.

Lukashin, A. V. and M. Borodovsky. 1998. GeneMark. hmm: new solutions for gene finding. Nucleic Acids Research 26: 1107-1115.

Nicorici D. and J. Astola. 2004. Segmentation of DNA into coding and noncoding regions based on recursive entropic segmentation and stop-codon statistics. EURASIP Journal on Applied Signal Processing, Special issue in Genomic Signal Processing 1: 81-91.

Parra G., P. Agarwal, J. F. Abril, T. Wiehe, J. W. Fickett and R. Guigó. 2003. Comparative gene prediction in human and mouse. Genome Research 13: 108-117.

Pettitt, A. N. 1980. A simple cumulative sum type statistic for the change-point problem with zero-one observations. Biometrika 67: 79-84.

Singer M. and P. Berg. 1993. Genes & Genomes. A Changing Perspective. University Science Books. Mill Valley, California. 929 p.

Stanke M., and S. Waack. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19 (Suppl. 2): ii215-ii225.

Ringo J. (2007). Genética Fundamental. Editorial Acribia. Zaragoza España. 398 p.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58: 267-288.