

Identificación de Regiones Codificantes de Proteínas Mediante la Transformada Modificada de Morlet

Jesús P. Mena-Chalco, Roberto M. Cesar-Jr.*
Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo
Rua do Matão 1010, São Paulo, SP, Brasil
{jmena,cesar}@vision.ime.usp.br

Abstract

An important topic in biological sequences analysis is gene identification, i.e., the identification of protein coding regions. In this context, several methods combine pattern recognition with knowledge collected from training datasets. Nonetheless, the problem of gene identification is far from being solved because the accuracy of these methods is still far from satisfactory. In the Msc. study, we introduced a new method, which does not need training dataset for identification of protein coding regions. This method is based on a new transform, here called Modified Morlet Transform, and defined because traditional time-scale transforms are not suitable to gene identification. We present the main obtained results thus showing that our method performs better than previous approaches based on the short time Fourier transform.

Keywords: gene identification, modified Morlet transform, digital signal processing, bioinformatics.

Resumen

Un tópico importante en el análisis de secuencias biológicas es la identificación de genes, i.e., la identificación de regiones codificantes de proteínas. En este contexto, diversos métodos combinan técnicas de reconocimiento de patrones con conocimiento obtenido de conjuntos de datos genómicos. Sin embargo, el problema de identificación de genes está en abierto porque la exactitud de esos métodos está aún distante de lo satisfactorio. En el estudio de maestría propusimos un nuevo método, que no necesita de conjuntos de entrenamiento, para la identificación de regiones codificantes. Este método se basa en una nueva transformada, denominada Transformada Modificada de Morlet y definida porque transformadas tiempo-escala tradicionales no son completamente apropiadas para la identificación de genes. Aquí presentamos los principales resultados obtenidos, los cuales muestran que nuestro método tiene mejor desempeño que los otros previos basados en la transformada de Fourier de tiempo reducido.

Palabras claves: identificación de genes, transformada modificada de Morlet, procesamiento digital de señales, bioinformática.

1. Introducción

El área de Bioinformática recientemente está recibiendo mucha importancia por la ayuda que brinda en el descubrimiento genómico y en el mejor entendimiento de los organismos biológicos. Cuando un nuevo organismo es secuenciado se desea obtener toda la información posible de su genoma, siendo un paso fundamental la identificación de genes presentes en su estructura genómica. Esta identificación corresponde a la determinación de las regiones codificantes de proteínas (CDS, *Coding Sequences*).

Las CDS típicamente presentan una organización periódica imperfecta de tres bases, la cual generalmente no está presente en las regiones intergénicas e intrones. En los últimos años, esa característica independiente

*Profesor orientador de la tesis de maestría.

de las especies fue analizada para poder explicar su origen [11, 18, 21, 22, 27] y así cuantificarla [10, 19]. En la literatura es comúnmente denominada de periodicidad de tres bases (TBP, *Three-Base Periodicity*), habiendo sido observada de manera similar para di-nucleótidos en cromosomas de bacterias [16]. Se cree que esa periodicidad muestra relaciones entre posiciones de nucleótidos en las CDS causada por la asimetría en la composición de bases respecto a las tres posiciones codificantes [11]. Pueden ser encontradas algunas excepciones de esta característica en CDS de secuencias de virus y mitocondrias [13]. En regiones intergénicas de *E. coli* fue encontrada periodicidad de aproximadamente 11 bases [12] y se sugiere que esta sea una propiedad típica de esas regiones, posiblemente, para la regulación de su transcripción genómica.

Nuevos métodos que combinan técnicas de reconocimiento de patrones y de procesamiento digital de señales (DSP, *Digital Signal Processing*) aplicados en Bioinformática fueron utilizados [7, 15, 26] por su robustez matemática y rapidez computacional. Los métodos de DSP fueron usados para identificar exones en células de eucariotos [2, 20, 23], mostrando resultados prometedores. Estos métodos, enfocados solamente en la búsqueda de regiones con TBP, son recomendables para uso en genomas en los cuales no exista conjuntos de secuencias de entrenamiento.

Los métodos de DSP para la identificación de CDS basados en la transformada de Fourier [2, 20] y en filtros digitales [23] no presentan formulaciones robustas debido a su dependencia en relación al tamaño de ventana para el análisis local, así como a la carencia de delimitaciones entre CDS y no CDS¹. La definición previa del tamaño de ventana es crítica [14] y para reducir tal dependencia, métodos alternativos que exploran diferentes tamaños de ventana, como la transformada en *wavelets*, fueron recientemente estudiados.

Una manera natural para la identificación de CDS mediante técnicas de multiresolución consiste en el uso de pequeñas o grandes escalas en cortas o largas CDS, respectivamente. De ese modo, las CDS cortas podrán ser analizadas por funciones de soporte pequeño y las CDS largas por funciones de soporte grande. En este contexto, la transformada en *wavelets* es un enfoque lógico para analizar las secuencias de ADN. Sin embargo, no es completamente apropiado porque la frecuencia de las funciones de análisis varían con la alteración de los valores de escala.

La principal contribución de la tesis de maestría [17] es la propuesta de un método para la identificación de CDS basado en la transformada modificada de Morlet (MMT, *Modified Morlet Transform*) que resuelve apropiadamente el problema de análisis de señales con frecuencia específica y de escala variable. La MMT fue igualmente introducida en la tesis. Adicionalmente, evaluamos el desempeño del método comparándolo con otro basado en la transformada de Fourier de tiempo reducido (STFT, *Short Time Fourier Transform*).

Este artículo está organizado de la siguiente manera: en la Sección 2 introducimos la MMT; en la Sección 3 describimos el conjunto de secuencias usado y detallamos nuestro método para la identificación de CDS así como las medidas usadas para evaluar su desempeño; finalmente presentamos los resultados y discutimos el desempeño del método en las secciones 4 y 5, respectivamente.

2. Transformada Modificada de Morlet

Una transformada multiescala de una señal u puede ser calculada por,

$$U(b, a) = \int u(x)f(x, b, a)dx, \quad (1)$$

donde $a > 0$ es el parámetro de escala, b el parámetro de espacio, y f la función de análisis.

En la Ecuación (1) diferentes funciones de análisis pueden ser adoptadas para transformar la señal u . En particular, funciones bien localizadas en el dominio de la frecuencia, como la función de Gabor (Gaussiana modulada) definida como [8],

$$g(x, a) = e^{-\frac{x^2}{2}} e^{j a x}, \quad (2)$$

y la función de Morlet [6],

$$\psi_M(x) = e^{-\frac{x^2}{2}} e^{j \omega_0 x}, \quad (3)$$

donde ω_0 es la frecuencia básica de ψ , son utilizadas para analizar señales de forma local y con diferentes frecuencias. Estas funciones no son completamente apropiadas para la identificación de CDS porque varían

¹Actualmente los métodos para la identificación de CDS basados en DSP muestran solamente una verificación visual de los resultados, presentando medidas cualitativas en la identificación, las cuales dificultan mucho una posible comparación con otros métodos.

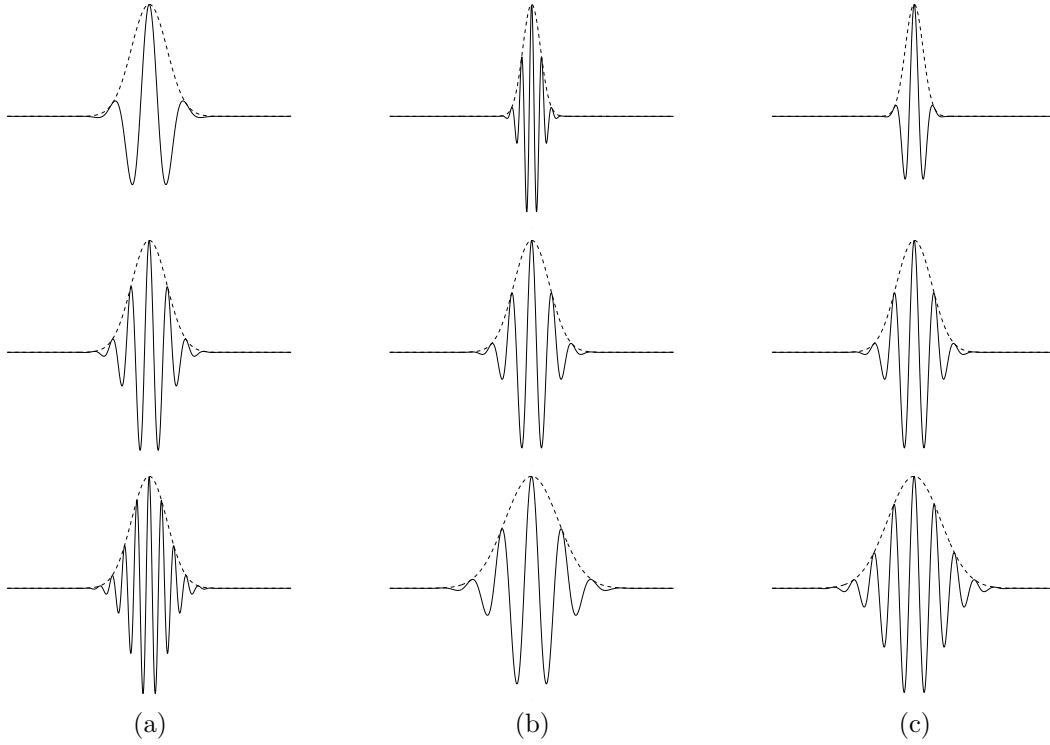


Figura 1: Representación de las funciones de análisis de (a) Gabor, donde la frecuencia de la exponencial compleja es variada, manteniendo constante la desviación estándar de la Gaussiana, (b) Morlet, donde la desviación estándar de la Gaussiana es variada así como la frecuencia de la exponencial compleja, y (c) Morlet modificado, donde la desviación estándar de la Gaussiana es variada, manteniendo constante la frecuencia de la exponencial compleja.

sus frecuencias con la alteración de la escala, i.e., la variación de la desviación estándar de la Gaussiana involucrada en la ecuación. En la Figura 1a-b, mostramos las diferencias entre ambas funciones donde la frecuencia de la exponencial compleja es variada.

Definimos una modificación de la función de Morlet para analizar localmente señales en una frecuencia específica y con escala variable. En la función de análisis de Morlet usamos el parámetro de escala a para mantener constante la frecuencia de la exponencial compleja, variando la desviación estándar de la Gaussiana (i.e., la escala),

$$U(b, a) = \int u(x) e^{-\frac{(x-b)^2}{2a^2}} e^{j\omega_0(x-b)} dx \quad (4)$$

Por lo tanto, la función de análisis ψ_{MM} de la Transformada Modificada de Morlet está definida por:

$$\psi_{MM}(x, a) = e^{-\frac{x^2}{2a^2}} e^{j\omega_0 x} \quad (5)$$

En la Figura 1 gráficamente ilustramos las diferencias entre las funciones de análisis de Gabor, Morlet y la modificación de Morlet.

3. Material y Método

3.1. Conjunto de Secuencias

Enfocamos nuestro estudio en el análisis de secuencias de ADN sintéticas y reales. Para fines de demostración del método usamos una secuencia de ADN de 8000 bp de *Caenorhabditis elegans* el cual contiene al gen F56F11.4 constituido por cinco exones (GenBank², número de acceso AF099922). Adicionalmente, usamos

²Disponible en <http://www.ncbi.nlm.nih.gov/Genbank/>.

varios conjuntos de secuencias de ADN. En el presente resumen tratamos solamente a un conjunto de 570 secuencias de vertebrados [4], con sus respectivos límites entre exones e intrones, y dos subconjuntos de este donde fueron ignoradas secuencias con tamaños de exones menores a 30 bp y 100 bp (vea en el Apéndice A algunas estadísticas de este conjunto). En el análisis de resultados para el caso de organismos eucariotes consideramos que en cada secuencias de ADN existe únicamente un gen. De tal manera que la primera y el última región identificada corresponderan respectivamente a la primera y última CDS.

3.2. Método para la Identificación de CDS

El método propuesto para la identificación de regiones codificantes de proteínas (Figura 2) está dividido en cuatro procesos: (1) mapeamiento numérico de secuencias de ADN en cuatro secuencias binarias, (2) aplicación de la MMT en cada secuencia binaria, (3) proyección de los espectros de las secuencias, y (4) *Thresholding* de los coeficientes de proyección para el establecimiento de los posibles límites entre CDS.

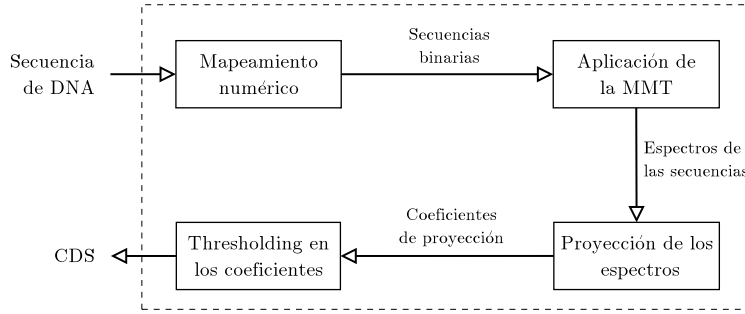


Figura 2: Diagrama de flujo de datos del método para la identificación de CDS (líneas punteadas). Cada bloque representa una operación (líneas sólidas) y cada flecha el flujo de datos entre procesos.

3.2.1. Mapeamiento Numérico de Nucleótidos

En este trabajo usamos las reglas 4–7 del mapeamiento fijo binario [3] para crear cuatro secuencias binarias donde cada una represente las posiciones de la adenina (A), citosina (C), guanina (G) y timina (T)³. Las reglas 1–3 no producen información relevante para nuestro método ya que la TBP es particularmente perdida.

Considerando una secuencia de ADN s , denotamos por u_A , u_C , u_G y u_T a las secuencias correspondientes a las cuatro reglas asociadas a los nucleótidos A, C, G y T, respectivamente. Esta representación redundante es preferible porque no depende de alguna atribución numérica adoptada en particular y ninguna estructura armónica de significancia biológica relevante es oculta o expuesta [1]. Una clasificación de esquemas de mapeamiento es detallado en [5].

3.2.2. Aplicación de la MMT

La MMT con diferentes escalas a y frecuencia angular ω_0 , siendo un múltiplo de tres⁴, es calculada para todas las secuencias binarias utilizando ψ_{MM} de N puntos. Las transformadas de las secuencias u_A , u_C , u_G y u_T , correspondientes a la secuencia s , son dadas respectivamente por:

$$U_A(b, a) = \int u_A(x) \psi_{MM}^*(x - b, a) dx \quad (6)$$

$$U_C(b, a) = \int u_C(x) \psi_{MM}^*(x - b, a) dx \quad (7)$$

$$U_G(b, a) = \int u_G(x) \psi_{MM}^*(x - b, a) dx \quad (8)$$

$$U_T(b, a) = \int u_T(x) \psi_{MM}^*(x - b, a) dx \quad (9)$$

³Estas secuencias también son conocidas como secuencias indicadoras, las cuales indican la posición relativa de cada nucleótido en la secuencia de ADN [24].

⁴La definición de $\omega_0 = N/3$ implica que la frecuencia angular de ψ_{MM} es un múltiplo de tres.

De esa forma, transformadas con diferentes escalas pueden ser aplicadas para el análisis de secuencias de ADN. Los mejores resultados obtenidos fueron para escalas separadas exponencialmente entre 0,2 y 0,7. En este trabajo, el espectro de cada secuencia binaria es definido como el módulo al cuadrado de sus coeficientes después de ser aplicada la transformada:

$$m_A(b, a) = |U_A(b, a)|^2 \quad (10)$$

$$m_C(b, a) = |U_C(b, a)|^2 \quad (11)$$

$$m_G(b, a) = |U_G(b, a)|^2 \quad (12)$$

$$m_T(b, a) = |U_T(b, a)|^2 \quad (13)$$

Así el espectro total, que combina las contribuciones de todas las transformadas, es la suma de los espectros de cada secuencia binaria:

$$M(b, a) = m_A(b, a) + m_C(b, a) + m_G(b, a) + m_T(b, a) \quad (14)$$

Con la formulación del espectro total intentamos parcialmente representar la interacción y la dependencia existente entre nucleótidos en el genoma. Una vez que la secuencia es representada mediante su espectro multiescala, diferentes técnicas pueden ser usadas para extraer información de ellas [8].

3.2.3. Proyección de los Espectros de las Secuencias

El espectro total de la secuencia analizada es proyectado en el eje de las posiciones para obtener una medida local de las regiones con TBP. Dada una secuencia de tamaño N , los coeficientes de proyección del espectro total aquí son definidos como:

$$M(b) = \sum_a M(b, a), \quad b = 0, \dots, N - 1 \quad (15)$$

Es importante destacar que estos coeficientes resultantes permiten comparar nuestro método usando la MMT y la STFT [20]. Por otro lado, la proyección en el eje de las escalas revela cual escala mantiene mas energía en la secuencia através de las posiciones:

$$M(a) = \sum_{b=1}^N M(b, a), \quad \forall a \quad (16)$$

3.2.4. Thresholding en los Coeficientes de Proyección

Una manera natural de localizar los límites entre CDS y no CDS es mediante la incorporación de un *threshold* en los coeficientes de proyección. *Thresholding* sobre M (Ecuación 15) permite excluir posiciones donde los coeficientes sean pequeños, i.e., todos los coeficientes menores que un valor dado son substituidos por cero [9]. En general, regiones con poca o ninguna TBP tienen coeficientes de proyección pequeños. Así, los coeficientes diferentes de cero serán aquellos asociados a las posibles CDS de la secuencia s .

3.3. Medidas de Desempeño

Fueron consideradas como medidas de exactitud de identificación la sensibilidad, especificidad y correlación aproximada, principalmente para comparar nuestro método usando la MMT y la STFT. Las medidas de exactitud fueron calculadas mediante el conteo de nucleótidos correspondientes a los verdaderos positivos (TP, *true positives*), verdaderos negativos (TN, *true negatives*), falsos positivos (FP, *false positives*) y falsos negativos (FN, *false negatives*) de las regiones identificadas con las verdaderas regiones codificantes [4].

Usualmente la sensibilidad es definida por:

$$Sn = \frac{TP}{TP + FN} \quad (17)$$

que representa la proporción de nucleótidos codificantes que son correctamente identificados como parte de las CDS. La especificidad es definida como:

$$Sp = \frac{TP}{TP + FP} \quad (18)$$

que representa la proporción de nucleótidos identificados como codificantes que son actualmente considerados como parte de las CDS.

En este trabajo, usamos la correlación aproximada como medida total de exactitud en la identificación de CDS,

$$AC = \frac{1}{2} \left[\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FN} + \frac{TN}{TN + FP} \right] - 1 \quad (19)$$

Estas medidas son comúnmente usadas para indicar el desempeño en las identificaciones realizadas computacionalmente [4]. Medidas subjetivas y otros componentes de desempeño no fueron considerados en el análisis de los resultados.

4. Resultados

Diferentes experimentos fueron realizados para evaluar la eficiencia del método propuesto usando la MMT y la STFT [20] en el análisis de secuencias sintéticas y reales. Con la MMT usamos una señal de 1200 puntos y 40 escalas separadas exponencialmente entre 0,2 y 0,7. Por otro lado, con la STFT usamos tamaños de ventana de 200, 300 y 400 puntos correspondientes a los valores aproximados del promedio y la desviación estándar de los exones pertenecientes al conjunto de secuencias usado (Cuadro 1 del Apéndice A).

Para fines de comparación, usamos *thresholds* en el intervalo de 5 y 95 en los coeficientes de proyección, cuando es considerada la MMT, y la suma de los coeficientes normalizados en la frecuencia tres, cuando es considerada la STFT. Medidas de sensibilidad, especificidad y correlación aproximada fueron estimadas para los diferentes valores de *threshold*.

4.1. Resultados Obtenidos de un Gene en Particular

Fue analizado el gen F56F11.4 de *C. elegans* con el método propuesto. Los espectrogramas de las secuencias binarias están mostradas en la Figura 3a. Los coeficientes mayores están representados en *hot colors*. En las Figuras 3b-d mostramos representaciones de los coeficientes asociados a cada secuencia binaria procesada. La unión indica la medida total, i.e., la suma de todos los coeficientes después de aplicar la transformada sobre cada secuencia. Los picos corresponden a las regiones donde la TBP está presente.

Observemos que los coeficientes correspondientes a las CDS depende del tamaño de ventana e intervalo de escalas usado en el análisis de secuencias de ADN⁵. El primer exón de 112 bp tiene débil TBP. Por lo tanto, métodos basados únicamente en búsqueda por TBP no serán capaces de identificarlo.

El uso de la MMT muestra una máxima especificidad de 0,90 en una sensibilidad de 0,88 (Figura 3e y Cuadro 2 del Apéndice B). Usando un *threshold* de 85 % obtuvimos una exactitud de 0,87. Sin embargo, usando la STFT con los tamaños de ventana considerados fue obtenida una máxima exactitud de 0,74 (Figura 3f). Es importante ver que el desempeño usando la STFT con tamaños de ventana de 300 puntos es mejor que la de 200 y 400 puntos. De esa forma, podemos afirmar que para tamaños de ventana fuera de ese intervalo las CDS podrían no ser identificadas apropiadamente.

4.2. Resultados Obtenidos de un Conjunto de Secuencias

En la Figura 4 mostramos las medidas de desempeño calculadas para el conjunto original y dos subconjuntos de este donde secuencias con tamaños de exones menores a 30 bp e 100 bp no fueron consideradas. Para el conjunto original usando la MMT, se obtuvo una máxima especificidad de 0,44 en una sensibilidad de 0,78 (Cuadro 3 del Apéndice B). Con un *threshold* de 80 % se obtuvo una máxima exactitud de 49 %. Por el contrario, usando la STFT con tamaño de ventana de 300 puntos, se obtuvo una máxima especificidad de 0,38 en una sensibilidad de 0,75. Con el mismo *threshold* de 80 % se obtuvo una máxima exactitud de 40 %.

En los otros dos subconjuntos, se presenta un comportamiento similar, donde la MMT tiene desempeño superior a la STFT (Cuadros 4 y 5 del Apéndice B). En el subconjunto cuyas secuencias contienen exones de tamaños mayores a 100 bp se obtuvo una máxima exactitud de 56 %.

⁵La definición de estos valores críticos son sumamente importantes para trabajos futuros.

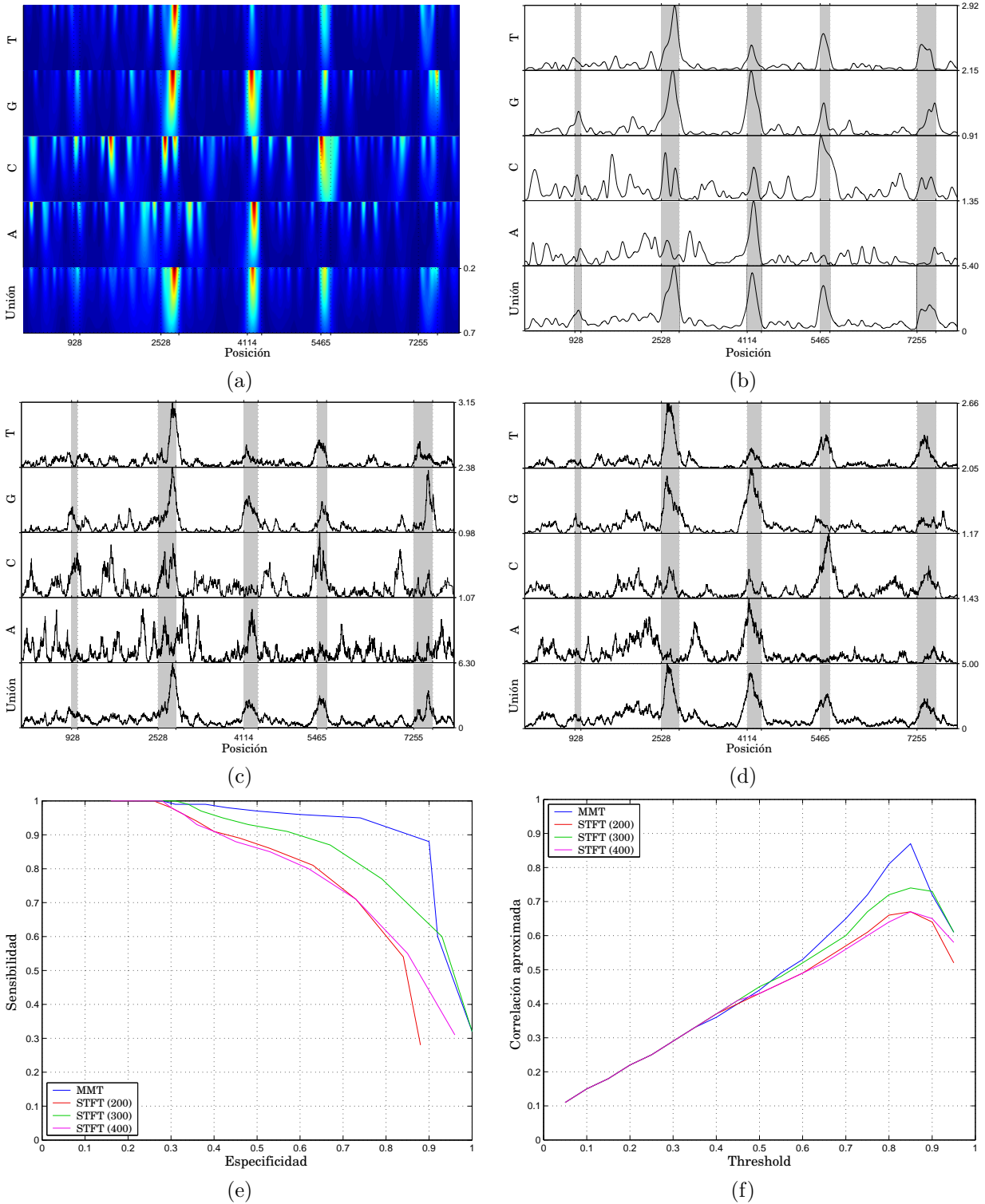


Figura 3: Resultado de la identificación de CDS en el gen F56F11.4. Usando la MMT: (a) los espectrogramas de las secuencias y (b) proyecciones de los espectros de las secuencias. Usando la STFT: los coeficientes normalizados con tamaños de ventana de (c) 200 y (d) 400 puntos. Las líneas verticalmente punteadas y las regiones sombreadas indican las posiciones de las regiones actualmente conocidas como codificantes. El desempeño en términos de (e) especificidad y sensibilidad, y (f) *threshold* y correlación aproximada también fueron calculados.

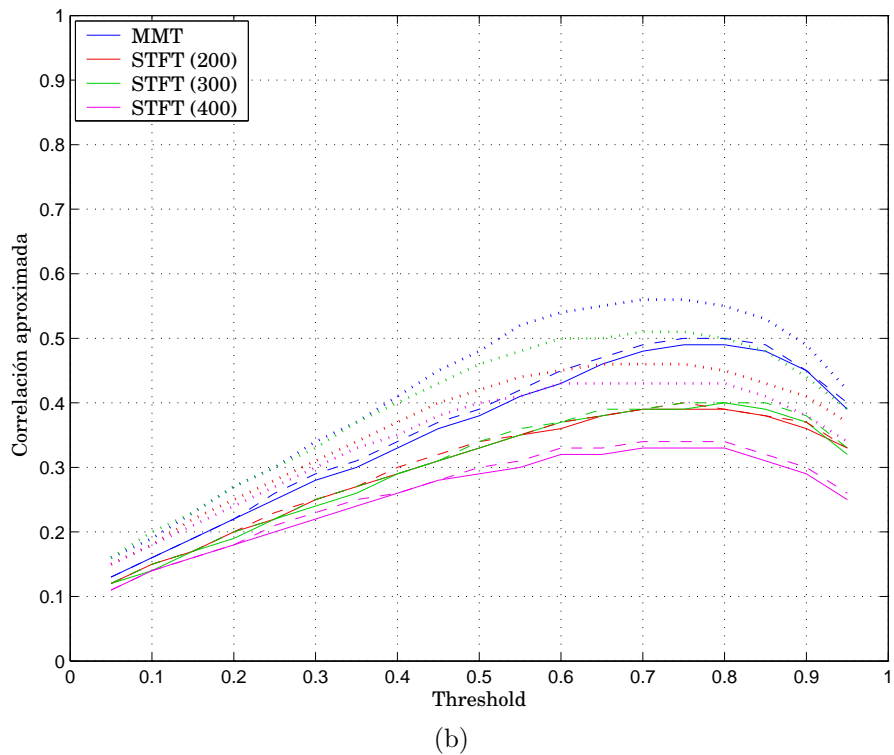
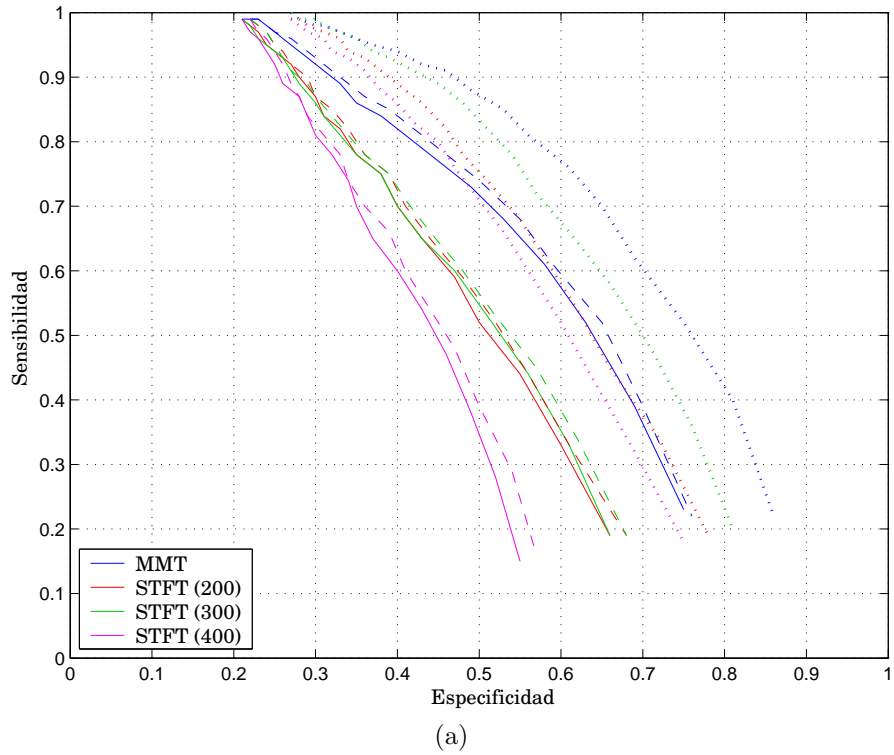


Figura 4: Desempeño en términos de (a) especificidad y sensibilidad, y (b) *threshold* y correlación aproximada de la identificación de CDS del método propuesto usando la MMT. Los tamaños de ventana usados con la STFT fueron de 200, 300 y 400 puntos. El desempeño correspondiente al conjunto original está mostrado con líneas sólidas. Los subconjuntos cuyas secuencias contienen exones mayores de 30 bp y 100 bp están mostrados con líneas discontinuas y punteadas, respectivamente.

5. Discusión

Los mejores desempeños fueron obtenidos, con un *threshold* de 80 %, en conjuntos cuyas secuencias contienen exones de tamaños superiores a 100 bp. Eso sugiere que nuestro método es adecuado para analizar secuencias cuyos tamaños de los exones sean mayores a 100 bp. Es interesante observar que el valor del *threshold* próximo de 85 % está relacionado con el 15 % de los nucleótidos pertenecientes a las CDS en el conjunto de secuencias usado (Cuadro 1 del Apéndice A). Creemos que valores de *threshold* adecuados podrían ser obtenidos de estadísticas de organismos taxonomicamente similares a los analizados.

En la Figura 5 mostramos los histogramas de los tamaños de los exones e intrones del conjunto de secuencias original posteriormente identificado con el método propuesto y un *threshold* de 80 %. Estas distribuciones en las frecuencia de sus tamaños mantienen la misma forma que las distribuciones calculadas para las secuencias del conjunto original (Figura 7 del Apéndice A). Esto es una buena señal del desempeño del método aquí tratado, y creemos que un análisis más detallado debe de seguir ese enfoque.

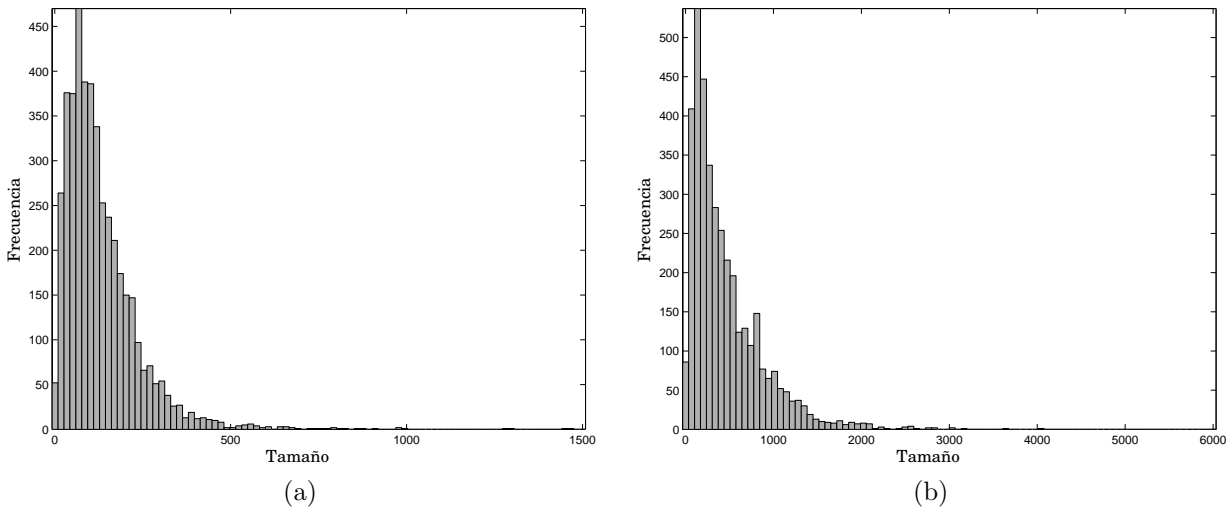


Figura 5: Histogramas de los tamaños de los (a) exones e (b) intrones para el conjunto de secuencias posteriormente identificado con nuestro método y con un *threshold* de 80 %.

Si bien los resultados obtenidos de nuestro método muestran una clara ventaja comparativa respecto de los otros, vemos que la exactitud obtenida por los métodos de DSP para la identificación de CDS, aquí incluido el nuestro, está aún lejos de lo idealmente esperado. El nivel máximo de exactitud alcanzado (56 % para el subconjunto cuyas secuencias contienen exones de tamaños mayores a 100 bp) se debe a tres cuestiones importantes que deben ser consideradas en la identificación de CDS:

- Existencia de TBP no uniforme en las CDS.
Como el que está presente en el primer exón del gen HSDAO del *Homo sapiens* de 9903 bp con número de acceso X78212-GenBank (Figura 6a).
- Existencia de poca o ninguna TBP en las CDS.
Como el que está presente en 34 exones del gen GGVITIIG del *Gallus gallus* de 20343 bp con número de acceso X13607-GenBank (Figura 6b). Observemos que solamente una CDS tiene alta TBP.
- Existencia de TBP en las regiones no codificantes.
Como el que está presente en la segunda región no codificante del gen MMACLGNA del *Mus musculus* de 2882 bp con número de acceso Z24722-GenBank (Figura 6c). Observemos que muchas regiones consideradas como no codificantes presentan altos valores en sus coeficientes de proyección. Creemos que posiblemente esas regiones pertenezcan a exones no anotados o a subregiones de pseudo-genes⁶[25].

⁶Secuencia de ADN derivada originalmente de genes codificantes de proteínas que fueron perdiendo su función.

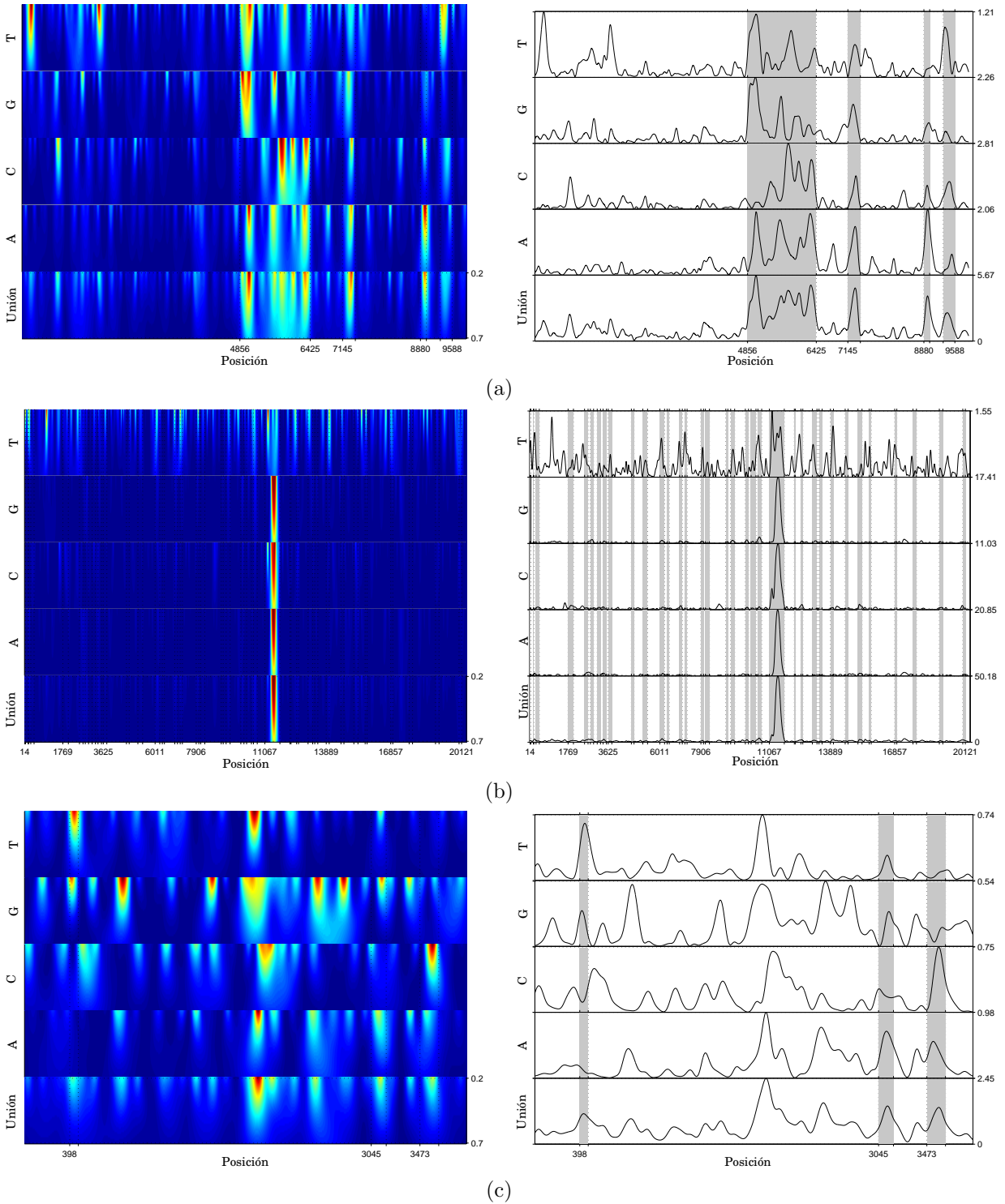


Figura 6: Espectrogramas (izquierda) y proyecciones de los espectros de las secuencias (derecha) calculadas para los genes (a) HSDAO del *Homo sapiens*, (b) GGVITIIG del *Gallus gallus*, y (c) MMACLGNA del *Mus musculus*. Las líneas verticalmente punteadas y las regiones sombreadas indican las posiciones de las regiones actualmente conocidas como codificantes.

6. Conclusiones

En la tesis de maestría [17] fue introducido un nuevo método para la identificación de regiones codificantes de proteínas, así como una modificación de la función de Morlet para el análisis de regiones periódicas. También fue evaluado comparativamente el desempeño del método con otro basado en la STFT.

El método propuesto en la tesis de maestría tiene un desempeño superior a los métodos previos basados en la STFT, porque el análisis multiresolución realizado permite examinar regiones cortas o largas con funciones de análisis periódicas y de soporte pequeño o grande, respectivamente. La característica significativa de este método es la robustez a las variaciones de escala. Actualmente, esa dependencia es una gran dificultad en los otros métodos del estado del arte.

Información Adicional

El texto completo de la tesis de maestría, los *scripts* usados y un sistema *on-line* para el análisis de secuencias de ADN mediante el método propuesto están disponibles en <http://www.vision.ime.usp.br/~jmena/DSPgenomics/>.

Agradecimientos

J. P. Mena-Chalco agradece a la CAPES (IEL Nacional – Brasil) por el apoyo financiero. R. M. Cesar-Jr. agradece a la FAPESP (2005/00587-5) y a la CNPq (300722/98-2, 474596/2004-4, 491323/2005-0). Agradecemos especialmente a la profesora Helaine Carrer por la orientación relativa a los aspectos biológicos. Deseamos extender nuestro agradecimiento a Yossi Zana y al profesor Luiz Velho por las discusiones y sugerencias en este trabajo.

Referencias

- [1] V. Afreixo, P. J. S. G. Ferreira, and D. Santos. Fourier analysis of symbolic data: A brief review. *Digital Signal Process.*, 14(6):523–530, 2004.
- [2] D. Anastassiou. Genomic signal processing. *IEEE Signal Processing Magazine*, 8(4):8–20, 2001.
- [3] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R.Ñ. Mantegna, M. E. Matsu, C.-K Peng, M. Simons, and H. E. Stanley. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Physical Review E*, 51(5):5084–5091, 1995.
- [4] M. Burset and R. Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34(3):353–367, 1996.
- [5] N. Chakravarthy, A. Spanias, L. D. Iasemidis, and K. Tsakalis. Autoregressive modelig and feature analysis of DNA sequences. *EURASIP Journal on Applied Signal Processing*, 1:13–28, 2004.
- [6] Y. T. Chan. *Wavelet Basics*. Kluwer Academic Publishers, Boston, 1995.
- [7] J. Chen, H. Li, K. Sun, and B. Kim. How will bioinformatics impact signal processing research? *IEEE Signal Processing Magazine*, 20(6):16–26, 2003.
- [8] L. F. Costa and R. M. Cesar Jr. *Shape Analysis and Classification: Theory and Practice*. CRC Press, Inc., Boca Raton, FL, USA, 2001.
- [9] D. L. Donoho, I. M. Johnstone, G. Kerkycharian, and D. Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society, B*, 57:301–337, 1995.
- [10] S. T. Eskesen, F.Ñ. Eskesen, B. Kinghorn, and A. Ruvinsky. Periodicity of DNA in exons. *Journal of Molecular Biology*, 5(12):1–11, 2004.
- [11] G. Gutierrez, J. L. Oliver, and A. Marin. On the origin of the periodicity of three in protein coding DNA sequences. *Journal of Theoretical Biology*, 167(4):413–414, 1994.

- [12] S. Hosid, E.Ñ. Trifonov, and A. Bolshoy. Sequence periodicity of *Escherichia coli* is concentrated in intergenic regions. *BMC Molecular Biology*, 5(14):1–7, 2004.
- [13] W. Li. The study of correlation structures of DNA sequences: A critical review. *Computers & Chemistry*, 21(4):257–271, 1997.
- [14] A. W.-C. Liew, H. Yan, and M. Yang. Pattern recognition techniques for the emerging field of bioinformatics: A review. *Pattern Recognition*, 38(11):2055–2073, 2005.
- [15] P. Liò. Wavelets in bioinformatics and computational biology: State of art and perspectives. *Bioinformatics*, 19(1):2–9, 2003.
- [16] I. Lopez-Villasenor, M. V. Jose, and J. Sanchez. Three-base periodicity patterns and self-similarity in whole bacterial chromosomes. *Biochemical and Biophysical Research Communications*, 325(2):467–478, 2004.
- [17] J. P. Mena-Chalco. Identificação de regiões codificantes de proteína através da transformada modificada de Morlet. Master’s thesis, Instituto de Matemática e Estatística – Universidade de São Paulo, October 2005.
- [18] B. Pierre, S. Brunak, Y. Chauviny, J. Engelbrecht, and A. Krogh. Periodic sequence patterns in human exons. *Proc Int Conf Intell Syst Mol Biol*, 3(3):30–38, 1995.
- [19] B. D. Silverman and R. Linsker. A measure of DNA periodicity. *Journal of Theoretical Biology*, 118(3):295–300, 1986.
- [20] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy. Prediction of probable genes by Fourier analysis of genomic sequences. *Bioinformatics*, 13(3):263–270, 1997.
- [21] E.Ñ. Trifonov and J. L. Sussman. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc Natl Acad Sci USA*, 77(7):3816–3820, 1980.
- [22] A. A. Tsonis, J. B. Elsner, and P. A. Tsonis. Periodicity in DNA coding sequences: Implications in gene evolution. *Journal of Theoretical Biology*, 151(3):323–331, 1991.
- [23] P. P. Vaidyanathan and B. Yoon. The role of signal-processing concepts in genomics and proteomics. *Journal of the Franklin Institute*, 341(1-2):111–135, 2004.
- [24] W. Wang and D. H. Johnson. Computing linear transforms of symbolic signals. *IEEE Transaction on Signal Processing*, 50(3):628–634, 2002.
- [25] M. Q. Zhang. Computational prediction of eukaryotic protein-coding genes. *Nature Reviews Genetics*, 3(9):698–709, 2002.
- [26] X. Zhang, F. Chen, Y. Zhang, S. C. Agner, M. Akay, Z. Lu, M. M. Y. Waye, and S. K. Tsui. Signal processing techniques in genomic engineering. *Proceedings of the IEEE*, 90(12):1822–1833, 2002.
- [27] V. B. Zhurkin. Periodicity in DNA primary structure is defined by secondary structure of the coded protein. *Nucleic Acid Research*, 9(8):1963–1971, 1981.

Apéndice A: Estadísticas del Conjunto de Secuencias Utilizado

El conjunto de secuencias utilizado pertenece a organismos eucariotas cuyos límites entre exones e intrones fueron anotados con base en las interpretaciones biológicas. Este conjunto está compuesto por 570 secuencias de ADN de vertebrados y fue presentado por M. Bursset y R. Guigó en el artículo [4]. Cada secuencia contiene exactamente un gen con por lo menos un intrón. Todas las secuencias consideradas son aquellas que codifican en productos de proteína completos. Para una descripción detallada de las secuencias consulte la página disponible en <http://www1.imim.es/databases/genomics96/index.html>.

Algunas estadísticas extraídas del conjunto de secuencias están presentadas en el Cuadro 1. Mostramos la cantidad de regiones de ADN, el número de bases en las regiones y el promedio y desviación estándar para cada tipo de región. En la Figura 7 están mostrados los histogramas de los tamaños de los exones e intrones de este conjunto.

Región	Cantidad	Bases	Tamaño		Contenido				
			Promedio	Desviación	A	C	G	T	N
Éxon	2649	444498(15,4 %)	168	222	108013	120970	124055	91459	1
Íntron	2079	1310452(45,3 %)	630	909	332478	298396	310586	368597	395
Inter-génica	1132	1137199(39,3 %)	1004	1464	296109	270067	267702	302844	477
Total	5860	2892149	—	—	736600	689433	702343	762900	873

Cuadro 1: Estadísticas extraídas del conjunto.

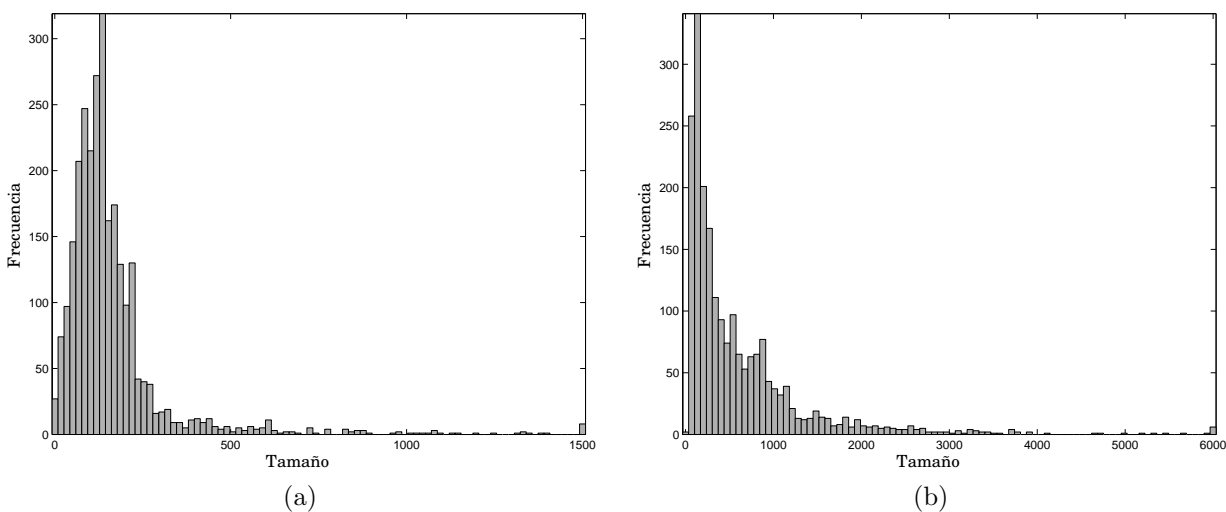


Figura 7: Histogramas de los tamaños de los (a) exones e (b) intrones para el conjunto de secuencias.

Apéndice B: Medidas de Desempeño Obtenidas

Mostramos en los Cuadros 2, 3, 4 y 5 las medidas de desempeño (Sección 3.3) calculadas para las secuencias de ADN consideradas en este trabajo. Fueron utilizadas la MMT con 40 escalas separadas exponencialmente entre 0,2 y 0,7, y la STFT con tamaños de ventana de 200, 300 y 400 puntos.

Threshold	MMT			STFT (200)			STFT (300)			STFT (400)		
	S_n	S_p	AC	S_n	S_p	AC	S_n	S_p	AC	S_n	S_p	AC
05 %	1,00	0,16	0,11	1,00	0,16	0,11	1,00	0,16	0,11	1,00	0,16	0,11
10 %	1,00	0,17	0,15	1,00	0,17	0,15	1,00	0,17	0,15	1,00	0,17	0,15
15 %	1,00	0,18	0,18	1,00	0,18	0,18	1,00	0,18	0,18	1,00	0,18	0,18
20 %	1,00	0,19	0,22	1,00	0,19	0,22	1,00	0,19	0,22	1,00	0,19	0,22
25 %	1,00	0,21	0,25	1,00	0,21	0,25	1,00	0,21	0,25	1,00	0,21	0,25
30 %	1,00	0,22	0,29	1,00	0,22	0,29	1,00	0,22	0,29	1,00	0,22	0,29
35 %	1,00	0,24	0,33	1,00	0,24	0,33	1,00	0,24	0,33	1,00	0,24	0,33
40 %	1,00	0,26	0,36	1,00	0,26	0,37	1,00	0,26	0,37	1,00	0,26	0,37
45 %	1,00	0,28	0,40	0,99	0,28	0,40	1,00	0,28	0,41	1,00	0,28	0,41
50 %	0,99	0,31	0,44	0,98	0,30	0,43	1,00	0,31	0,45	0,98	0,30	0,43
55 %	0,99	0,34	0,49	0,96	0,33	0,46	0,99	0,34	0,48	0,96	0,33	0,46
60 %	0,99	0,38	0,53	0,94	0,36	0,49	0,97	0,37	0,52	0,93	0,36	0,49
65 %	0,98	0,43	0,59	0,91	0,40	0,53	0,95	0,42	0,56	0,91	0,40	0,52
70 %	0,97	0,50	0,65	0,89	0,46	0,57	0,93	0,48	0,60	0,88	0,45	0,56
75 %	0,96	0,60	0,72	0,86	0,53	0,61	0,91	0,57	0,67	0,85	0,53	0,60
80 %	0,95	0,74	0,81	0,81	0,63	0,66	0,87	0,67	0,72	0,80	0,62	0,64
85 %	0,88	0,90	0,87	0,71	0,73	0,67	0,77	0,79	0,74	0,71	0,73	0,67
90 %	0,60	0,92	0,72	0,54	0,84	0,64	0,60	0,93	0,73	0,55	0,85	0,65
95 %	0,32	1,00	0,61	0,28	0,88	0,52	0,32	1,00	0,61	0,31	0,96	0,58

Cuadro 2: Medidas de desempeño estimadas para el gene F56F11.4.

Threshold	MMT			STFT (200)			STFT (300)			STFT (400)		
	S_n	S_p	AC	S_n	S_p	AC	S_n	S_p	AC	S_n	S_p	AC
05 %	0,99	0,21	0,13	0,99	0,21	0,12	0,99	0,21	0,12	0,99	0,21	0,11
10 %	0,99	0,23	0,16	0,98	0,22	0,15	0,98	0,22	0,14	0,97	0,22	0,14
15 %	0,98	0,24	0,19	0,97	0,23	0,17	0,96	0,23	0,17	0,96	0,23	0,16
20 %	0,97	0,25	0,22	0,95	0,24	0,20	0,95	0,24	0,19	0,94	0,24	0,18
25 %	0,96	0,26	0,25	0,93	0,26	0,22	0,93	0,26	0,22	0,92	0,25	0,20
30 %	0,94	0,28	0,28	0,92	0,27	0,25	0,91	0,27	0,24	0,89	0,26	0,22
35 %	0,93	0,29	0,30	0,90	0,28	0,27	0,89	0,28	0,26	0,87	0,28	0,24
40 %	0,91	0,31	0,33	0,87	0,30	0,29	0,86	0,30	0,29	0,84	0,29	0,26
45 %	0,89	0,33	0,36	0,84	0,31	0,31	0,84	0,31	0,31	0,81	0,30	0,28
50 %	0,86	0,35	0,38	0,82	0,33	0,33	0,81	0,33	0,33	0,78	0,32	0,29
55 %	0,84	0,38	0,41	0,78	0,35	0,35	0,78	0,35	0,35	0,74	0,34	0,30
60 %	0,81	0,41	0,43	0,75	0,38	0,36	0,75	0,38	0,37	0,70	0,35	0,32
65 %	0,78	0,44	0,46	0,70	0,40	0,38	0,70	0,40	0,38	0,65	0,37	0,32
70 %	0,73	0,49	0,48	0,65	0,43	0,39	0,65	0,43	0,39	0,60	0,40	0,33
75 %	0,68	0,53	0,49	0,59	0,47	0,39	0,60	0,47	0,39	0,54	0,43	0,33
80 %	0,61	0,58	0,49	0,52	0,50	0,39	0,53	0,51	0,40	0,47	0,46	0,33
85 %	0,52	0,63	0,48	0,44	0,55	0,38	0,44	0,56	0,39	0,38	0,49	0,31
90 %	0,39	0,69	0,45	0,33	0,60	0,36	0,33	0,61	0,37	0,28	0,52	0,29
95 %	0,23	0,75	0,39	0,19	0,66	0,33	0,19	0,66	0,32	0,15	0,55	0,25

Cuadro 3: Medidas de desempeño estimadas para el conjunto original.

Threshold	MMT			STFT (200)			STFT (300)			STFT (400)		
	S_n	S_p	AC	S_n	S_p	AC	S_n	S_p	AC	S_n	S_p	AC
05 %	0,99	0,22	0,13	0,99	0,22	0,12	0,99	0,22	0,12	0,99	0,22	0,11
10 %	0,99	0,23	0,16	0,98	0,23	0,15	0,98	0,23	0,15	0,97	0,23	0,14
15 %	0,98	0,24	0,19	0,97	0,24	0,17	0,97	0,24	0,17	0,96	0,24	0,16
20 %	0,97	0,25	0,22	0,95	0,25	0,20	0,95	0,25	0,20	0,94	0,25	0,18
25 %	0,96	0,27	0,26	0,94	0,26	0,23	0,93	0,26	0,22	0,92	0,26	0,21
30 %	0,95	0,28	0,29	0,92	0,27	0,25	0,91	0,27	0,25	0,89	0,27	0,23
35 %	0,93	0,30	0,31	0,90	0,29	0,27	0,89	0,29	0,27	0,87	0,28	0,25
40 %	0,91	0,32	0,34	0,87	0,30	0,30	0,87	0,30	0,29	0,84	0,29	0,26
45 %	0,89	0,34	0,37	0,85	0,32	0,32	0,84	0,32	0,31	0,81	0,31	0,28
50 %	0,87	0,36	0,39	0,82	0,34	0,34	0,81	0,34	0,34	0,78	0,33	0,30
55 %	0,85	0,39	0,42	0,78	0,36	0,35	0,78	0,36	0,36	0,74	0,34	0,31
60 %	0,82	0,42	0,45	0,75	0,39	0,37	0,75	0,39	0,37	0,70	0,36	0,33
65 %	0,78	0,46	0,47	0,70	0,41	0,38	0,71	0,41	0,39	0,66	0,39	0,33
70 %	0,74	0,50	0,49	0,65	0,44	0,39	0,66	0,44	0,39	0,60	0,41	0,34
75 %	0,68	0,55	0,50	0,59	0,48	0,40	0,60	0,48	0,40	0,54	0,44	0,34
80 %	0,61	0,59	0,50	0,52	0,52	0,39	0,53	0,52	0,40	0,48	0,47	0,34
85 %	0,52	0,65	0,49	0,44	0,56	0,38	0,45	0,57	0,40	0,39	0,50	0,32
90 %	0,39	0,70	0,45	0,33	0,61	0,37	0,34	0,62	0,38	0,29	0,54	0,30
95 %	0,22	0,76	0,40	0,19	0,68	0,33	0,19	0,68	0,33	0,16	0,57	0,26

Cuadro 4: Medidas de desempeño estimadas para el subconjunto cuyas secuencias contienen exones de tamaños mayores a 30 bp.

Threshold	MMT			STFT (200)			STFT (300)			STFT (400)		
	S_n	S_p	AC	S_n	S_p	AC	S_n	S_p	AC	S_n	S_p	AC
05 %	1,00	0,27	0,16	0,99	0,27	0,15	1,00	0,27	0,16	0,99	0,27	0,15
10 %	0,99	0,28	0,19	0,99	0,28	0,18	0,99	0,28	0,20	0,98	0,28	0,18
15 %	0,98	0,30	0,23	0,98	0,29	0,22	0,99	0,30	0,23	0,97	0,29	0,21
20 %	0,98	0,31	0,27	0,97	0,31	0,25	0,98	0,31	0,27	0,96	0,31	0,24
25 %	0,97	0,33	0,30	0,96	0,33	0,28	0,97	0,33	0,30	0,94	0,32	0,27
30 %	0,96	0,35	0,34	0,94	0,34	0,31	0,96	0,35	0,33	0,93	0,34	0,30
35 %	0,95	0,37	0,37	0,93	0,36	0,34	0,94	0,37	0,37	0,91	0,36	0,33
40 %	0,94	0,40	0,41	0,91	0,38	0,37	0,93	0,39	0,40	0,88	0,38	0,35
45 %	0,92	0,43	0,45	0,88	0,41	0,40	0,91	0,42	0,43	0,86	0,40	0,38
50 %	0,91	0,46	0,48	0,86	0,43	0,42	0,89	0,45	0,46	0,83	0,42	0,40
55 %	0,88	0,49	0,52	0,82	0,46	0,44	0,86	0,48	0,48	0,80	0,45	0,41
60 %	0,85	0,53	0,54	0,78	0,48	0,45	0,82	0,51	0,50	0,76	0,47	0,43
65 %	0,81	0,56	0,55	0,74	0,51	0,46	0,78	0,54	0,50	0,71	0,50	0,43
70 %	0,76	0,61	0,56	0,68	0,55	0,46	0,72	0,57	0,51	0,66	0,53	0,43
75 %	0,70	0,65	0,56	0,62	0,58	0,46	0,66	0,61	0,51	0,60	0,56	0,43
80 %	0,62	0,69	0,55	0,54	0,62	0,45	0,58	0,66	0,50	0,52	0,60	0,43
85 %	0,52	0,75	0,53	0,45	0,66	0,43	0,48	0,71	0,48	0,43	0,64	0,41
90 %	0,40	0,81	0,49	0,34	0,72	0,41	0,36	0,76	0,44	0,32	0,69	0,38
95 %	0,22	0,86	0,42	0,19	0,78	0,37	0,20	0,81	0,39	0,18	0,75	0,34

Cuadro 5: Medidas de desempeño estimadas para el subconjunto cuyas secuencias contienen exones de tamaños mayores a 100 bp.