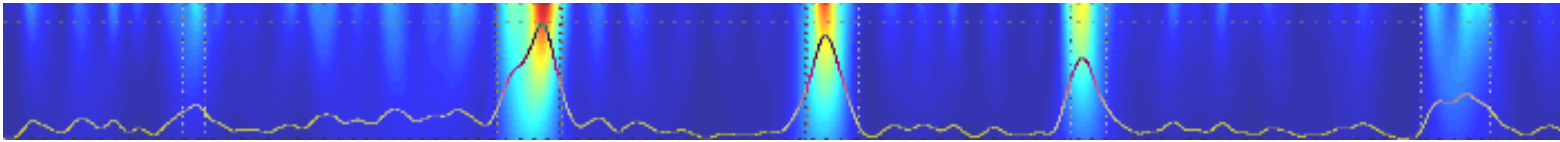


# Identificación de Regiones Codificantes de Proteínas Mediante la Transformada Modificada de Morlet



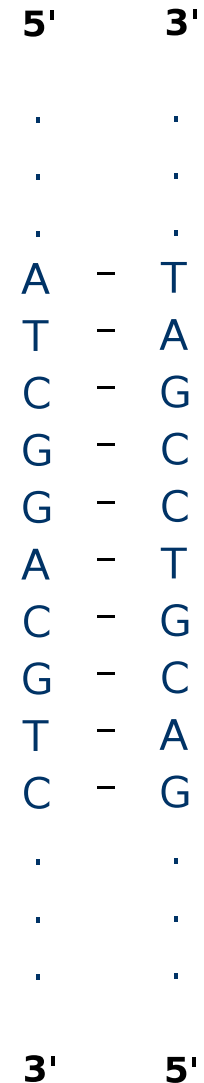
**Jesús P. Mena-Chalco**  
**Roberto Marcondes Cesar Jr.**

Departamento de Ciência da Computação  
Instituto de Matemática e Estatística  
Universidade de São Paulo

**XIII Concurso Latinoamericano de Tesis de Maestría**  
**32a Conferencia Latinoamericana de Informática**

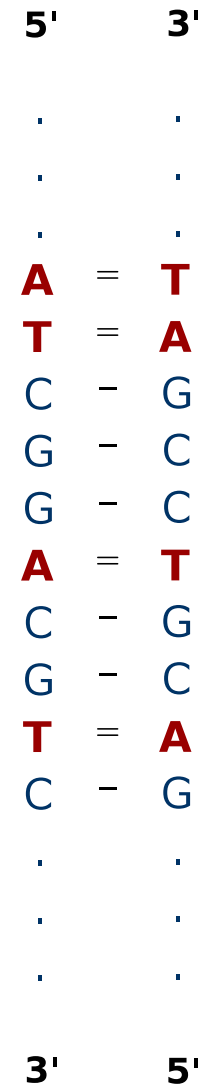
# ADN

- Todo organismo vivo almacena su información biológica en la forma de moléculas de ADN doblemente enlazadas y formadas por nucleótidos.
- El ADN es representado como una cinta doble, complementar e antiparalela.
- Son 4 los tipos de bases:
  - A: Adenina.
  - C: Citocina.
  - G: Guanina.
  - T: Timina.



# ADN

- Todo organismo vivo almacena su información biológica en la forma de moléculas de ADN doblemente enlazadas y formadas por nucleótidos.
- El ADN es representado como una cinta doble, complementar e antiparalela.
- Son 4 los tipos de bases:
  - A: Adenina.
  - C: Citocina.
  - G: Guanina.
  - T: Timina.

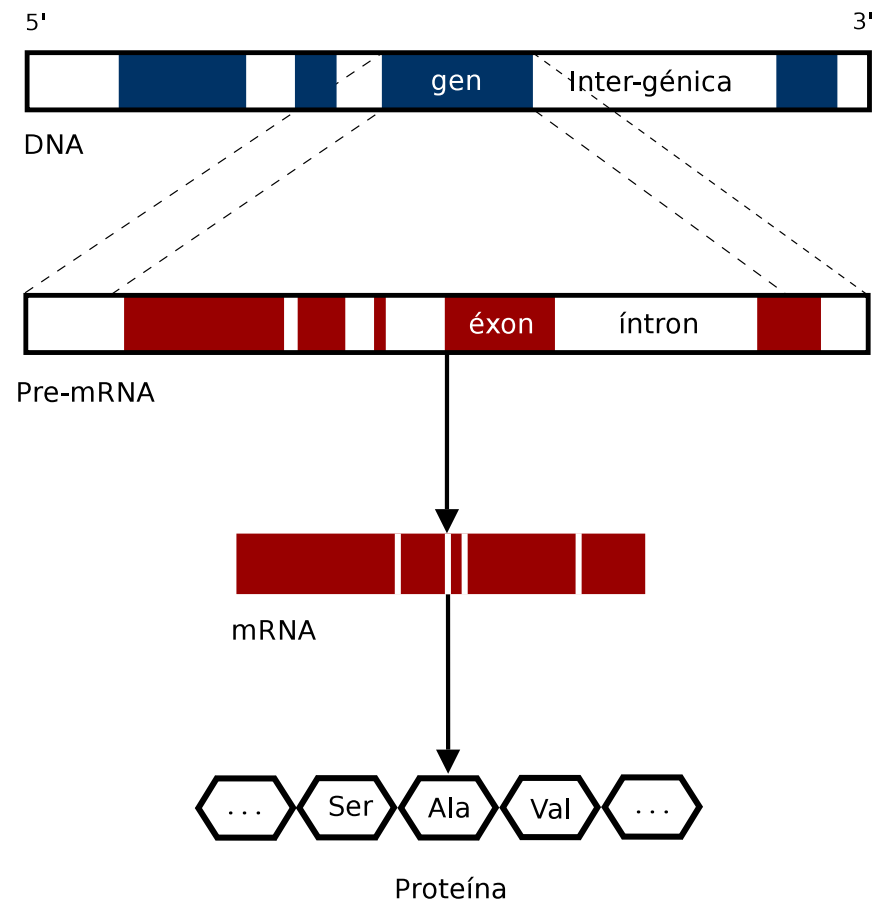


# Genes

Un gen es una región que expresa o controla una proteína.

Sub-regiones:

1. De reconocimiento (promotora);
2. De inicio de transcripción;
3. Región no-traducida 5';
4. De inicio de traducción (*start codon*);
5. **Región para la codificación de proteína (CDS);**
6. De traducción (*stop codon*);
7. Región no-traducida 3';
8. De poliadenilación (*polyA*, eucariotos);
9. De fin de transcripción.



# Contextualización

- Un tópico importante en el análisis de secuencias biológicas es la búsqueda de genes (identificación de regiones codificantes de proteína).
- Metodologías computacionales para identificar genes y otras regiones funcionales fueron desarrolladas en los últimos 20 años.
- Los métodos de procesamiento digital de señales (DSP) tiene un papel importante en ese contexto.
- Los métodos de DSP brindan una base robusta para la identificación de regiones codificantes de proteína.

# Estructura

1. El problema de la identificación de genes.
2. Métodos de DSP para la identificación de regiones codificantes.
3. Transformada modificada de Morlet.
4. Método propuesto.
5. Resultados.
6. Conclusiones.

# El problema: Identificación de Genes

Categorías que agrupan abordajes para su solución:

- Métodos basados en reconocimiento de patrones:
  - Búsqueda por sitios: se busca la presencia o ausencia de una secuencia específica, patrón o consenso asociado a la expresión génica;
  - **Búsqueda por contenido:** se busca segmentos con propiedades específicas.
- Métodos basados en comparaciones por homología con proteínas.
- Métodos basados en el uso de *expressed sequence tags* (ESTs).

# Periodicidad en las Regiones Codificantes

- Las regiones codificantes típicamente presentan una organización periódica imperfecta de tres bases, la cual generalmente no está presente en las regiones intergénicas e intrones.
- Esa característica, independiente de las especies, fue analizada para poder explicar su origen y así poder cuantificarla.
- En la literatura es comúnmente denominada de periodicidad de tres bases (TBP, *three-base periodicity*).
- Pueden ser encontradas algunas excepciones de esa característica en regiones codificantes de secuencias de virus y mitocondrias.

# Trabajos Relacionados

- Basados en la STFT [Tiwari et al., 1997; Anastassiou, 2001] y filtros digitales [Vaidyanathan & Yoon, 2004].
  - Dependencia del tamaño de ventana.
  - Dificultad en la determinación automática de límites entre regiones codificantes.

# Trabajos Relacionados

- Basados en la STFT [Tiwari et al., 1997; Anastassiou, 2001] y filtros digitales [Vaidyanathan & Yoon, 2004].
  - Dependencia del tamaño de ventana.
  - Dificultad en la determinación automática de límites entre regiones codificantes.
- Basados en *Wavelets* [Chen & Zhang, 2003; Ning et al., 2003].
  - Las frecuencias de las funciones de análisis varían con la alteración de los valores de escala.

# Transformada Multiescala

Una transformada multiescala de una señal  $u$  puede ser calculada como:

$$U(b, a) = \int u(x)\psi(x, b, a)dx$$

- $\psi$  función de análisis.
- $a > 0$  parámetro de escala.
- $b$  parámetro de espacio.

Diferentes funciones de análisis pueden ser adoptadas para transformar la señal  $u$ .

# Transformada de Fourier de Tiempo Reducido

En la STFT es usada una función de análisis de Gabor (Gaussiana modulada) que es bien localizada en el dominio de las frecuencias:

$$\psi_{\text{STFT}}(x, b, a) = e^{-\frac{(x-b)^2}{2}} e^{ja(x-b)}$$

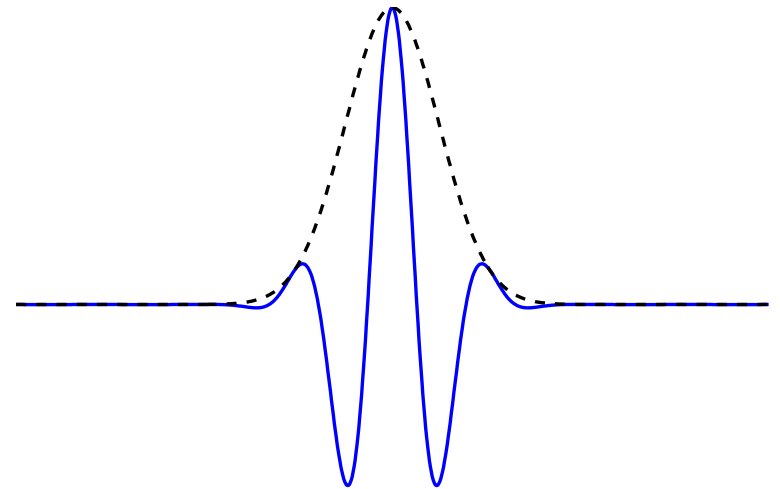
# Transformada de Fourier de Tiempo Reducido

En la STFT es usada una función de análisis de Gabor (Gaussiana modulada) que es bien localizada en el dominio de las frecuencias:

$$\psi_{\text{STFT}}(x, b, a) = e^{-\frac{(x-b)^2}{2}} e^{ja(x-b)}$$

La frecuencia de la exponencial compleja es variada, manteniendo constante la desviación estándar de la Gaussiana.

$$a = 3$$



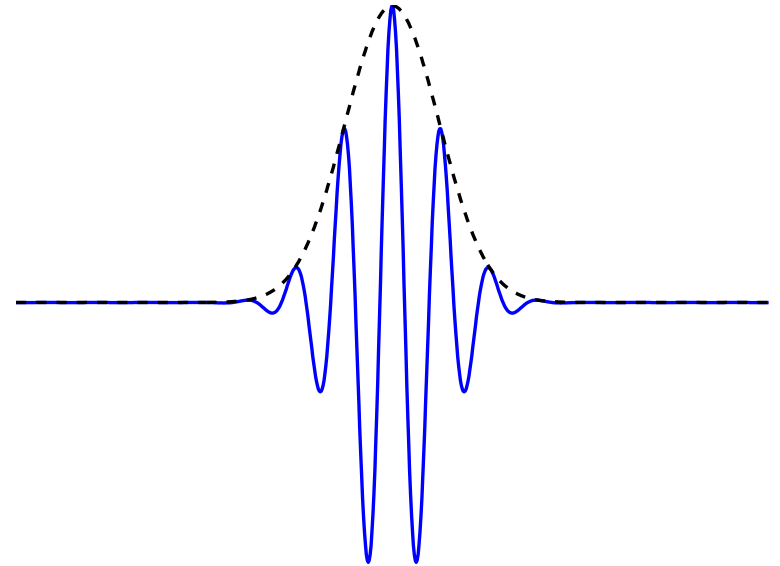
# Transformada de Fourier de Tiempo Reducido

En la STFT es usada una función de análisis de Gabor (Gaussiana modulada) que es bien localizada en el dominio de las frecuencias:

$$\psi_{\text{STFT}}(x, b, a) = e^{-\frac{(x-b)^2}{2}} e^{ja(x-b)}$$

La frecuencia de la exponencial compleja es variada, manteniendo constante la desviación estándar de la Gaussiana.

$$a = 6$$



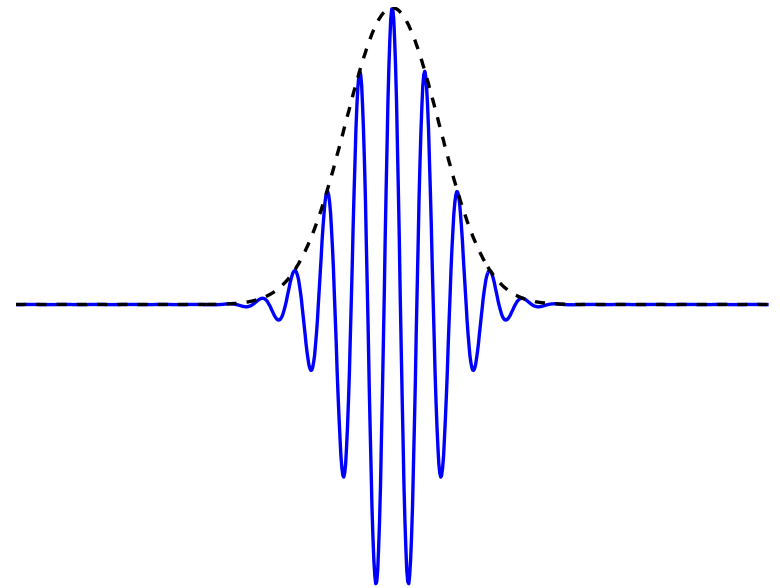
# Transformada de Fourier de Tiempo Reducido

En la STFT es usada una función de análisis de Gabor (Gaussiana modulada) que es bien localizada en el dominio de las frecuencias:

$$\psi_{\text{STFT}}(x, b, a) = e^{-\frac{(x-b)^2}{2}} e^{ja(x-b)}$$

La frecuencia de la exponencial compleja es variada, manteniendo constante la desviación estándar de la Gaussiana.

$$a = 9$$



# Transformada en *Wavelet* de Morlet

La función de análisis de Morlet es utilizada para analizar señales de forma local y con diferentes frecuencias:

$$\psi_{\text{MT}}(x, b, a) = e^{-\frac{(x-b)^2}{2}} e^{j\omega_0\left(\frac{x-b}{a}\right)}$$

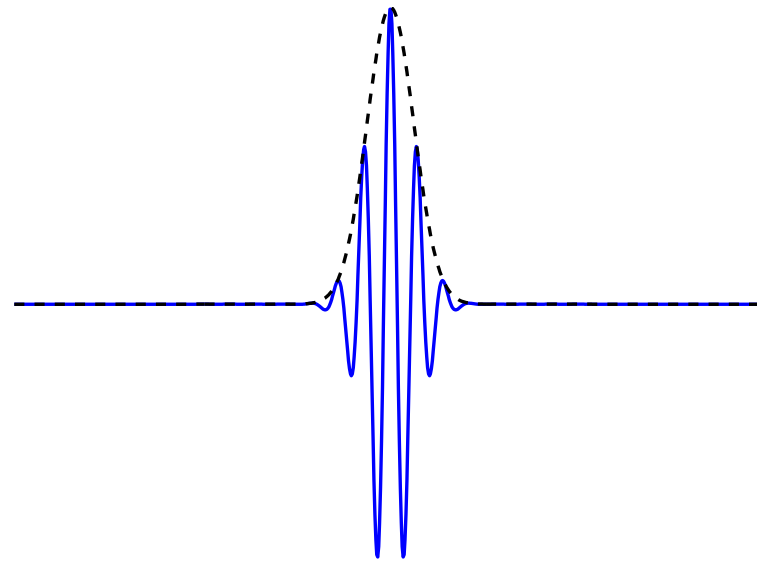
# Transformada en *Wavelet* de Morlet

La función de análisis de Morlet es utilizada para analizar señales de forma local y con diferentes frecuencias:

$$\psi_{\text{MT}}(x, b, a) = e^{-\frac{(x-b)^2}{2a}} e^{j\omega_0\left(\frac{x-b}{a}\right)}$$

La desviación estándar de la Gaussiana es variada así como la frecuencia de la exponencial compleja.

$$a = 0,5$$



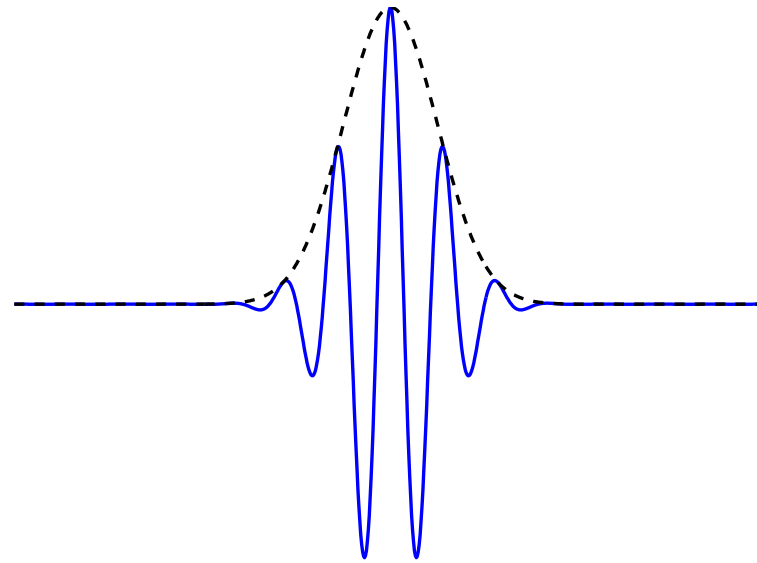
# Transformada en *Wavelet* de Morlet

La función de análisis de Morlet es utilizada para analizar señales de forma local y con diferentes frecuencias:

$$\psi_{\text{MT}}(x, b, a) = e^{-\frac{(x-b)^2}{2a^2}} e^{j\omega_0\left(\frac{x-b}{a}\right)}$$

La desviación estándar de la Gaussiana es variada así como la frecuencia de la exponencial compleja.

$$a = 1$$



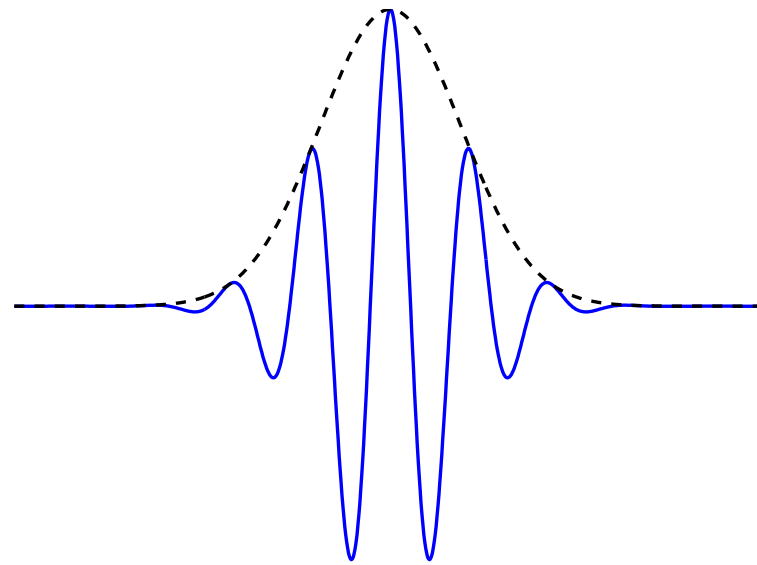
# Transformada en *Wavelet* de Morlet

La función de análisis de Morlet es utilizada para analizar señales de forma local y con diferentes frecuencias:

$$\psi_{\text{MT}}(x, b, a) = e^{-\frac{(x-b)^2}{2}} e^{j\omega_0\left(\frac{x-b}{a}\right)}$$

La desviación estándar de la Gaussiana es variada así como la frecuencia de la exponencial compleja.

$$a = 1,5$$



# Transformada Modificada de Morlet (MMT)

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.

$$\psi_{\text{MT}}(x, b, a) = e^{-\frac{(x-b)^2}{2a^2}} e^{j\omega_0 \frac{(x-b)}{a}}$$

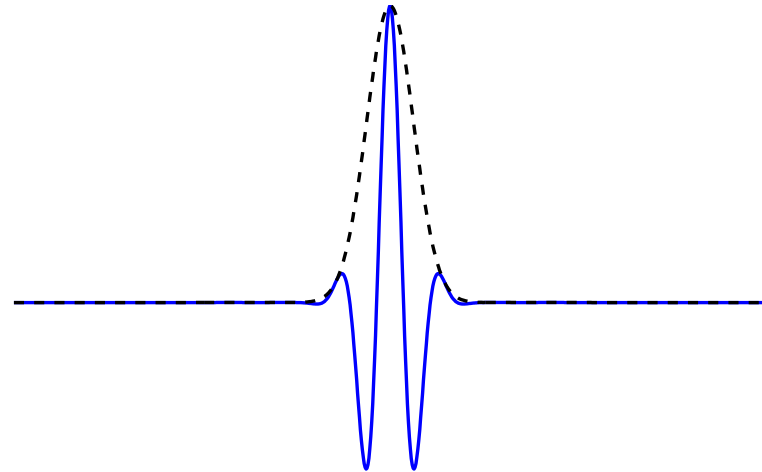
$$\psi_{\text{MMT}}(x, b, a) = e^{-\frac{(x-b)^2}{2a^2}} e^{j\omega_0(x-b)}$$

# Transformada Modificada de Morlet (MMT)

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

La desviación estándar de la Gaussiana es variada, manteniendo constante la frecuencia de la exponencial compleja.

$$a = 0,5$$

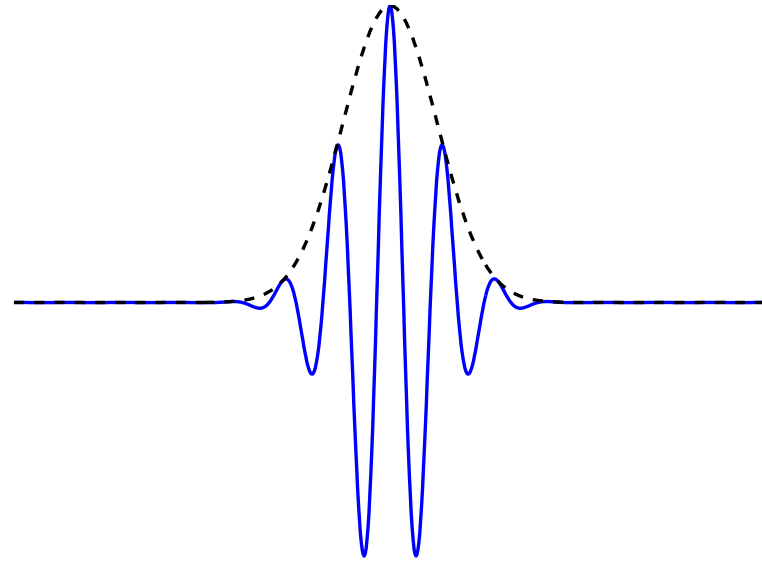


# Transformada Modificada de Morlet (MMT)

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

La desviación estándar de la Gaussiana es variada, manteniendo constante la frecuencia de la exponencial compleja.

$$a = 1$$

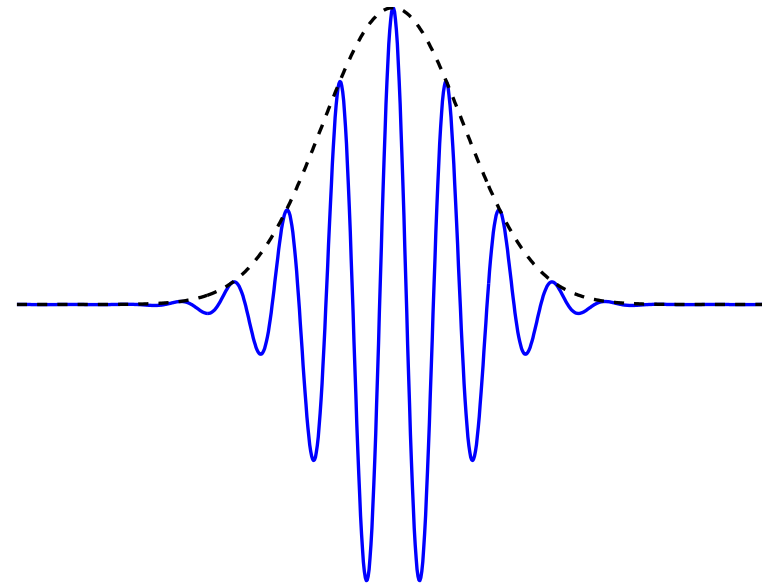


# Transformada Modificada de Morlet (MMT)

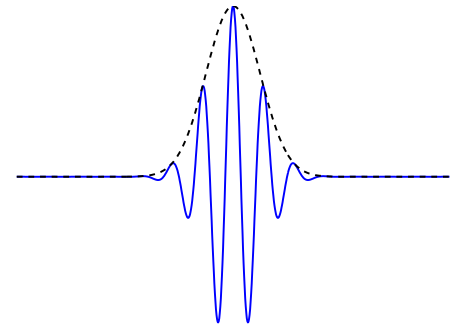
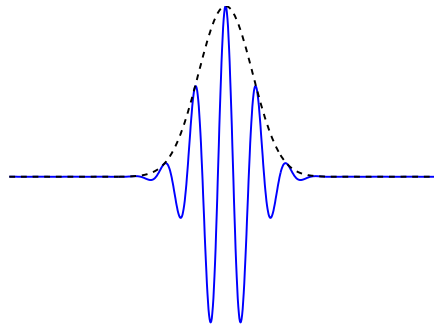
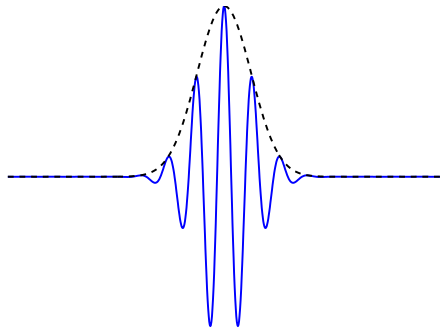
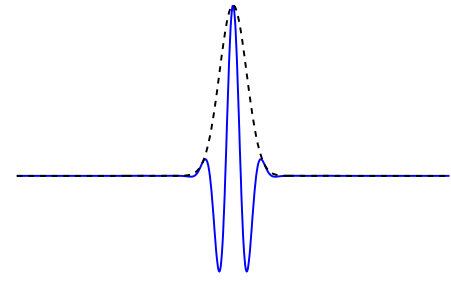
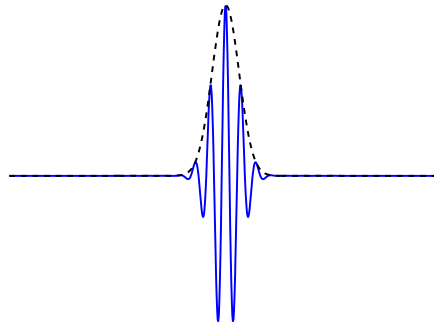
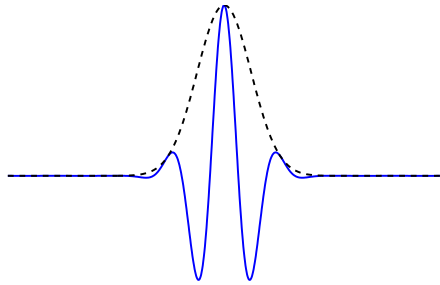
$$U(b, a) = \int u(x) e^{-\frac{(x-b)^2}{2a^2}} e^{j\omega_0(x-b)} dx$$

La desviación estándar de la Gaussiana es variada, manteniendo constante la frecuencia de la exponencial compleja.

$$a = 1,5$$



# Funciones de Análisis

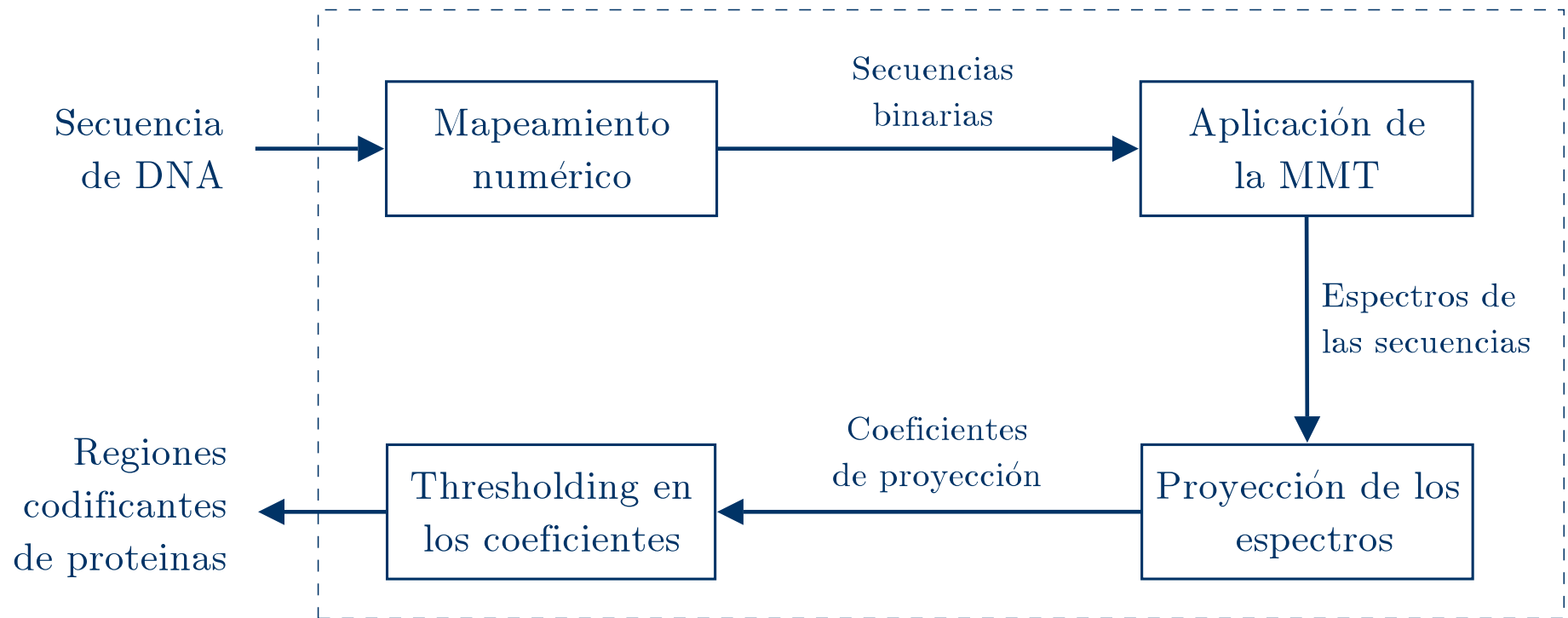


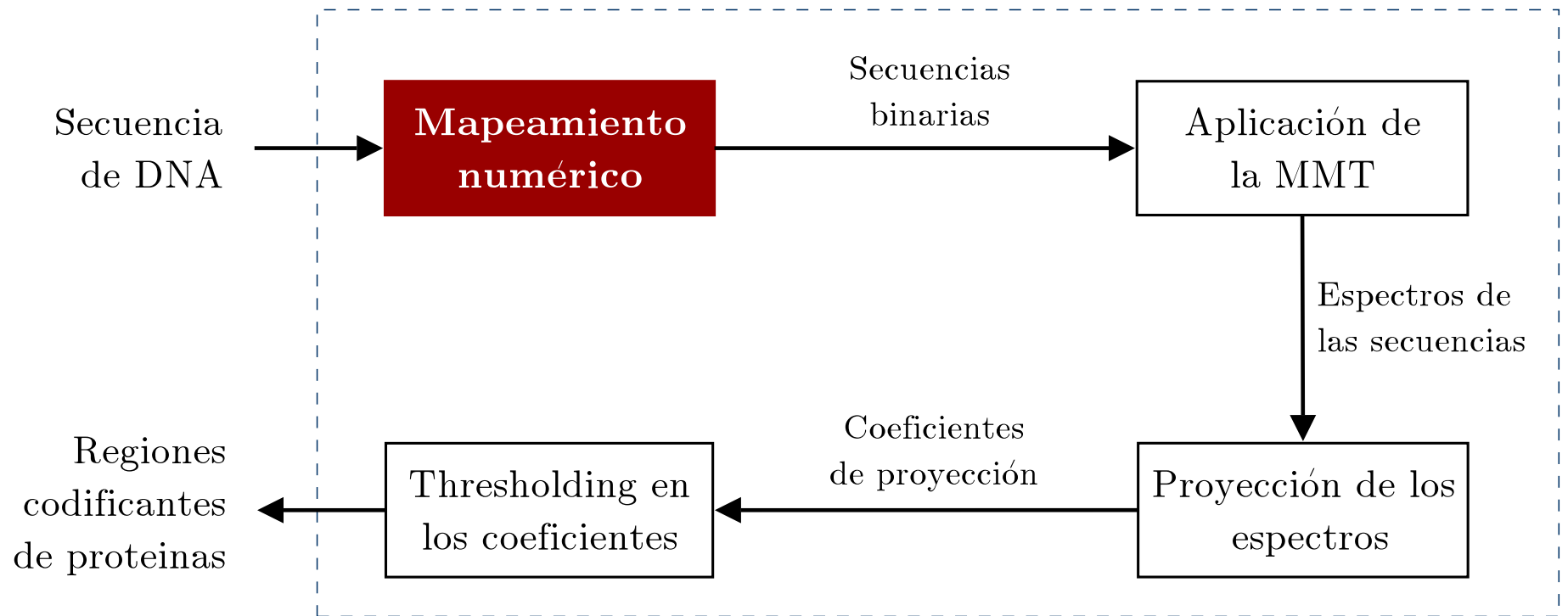
Gabor

Morlet

Morlet modificado

# Identificación de Regiones Codificantes





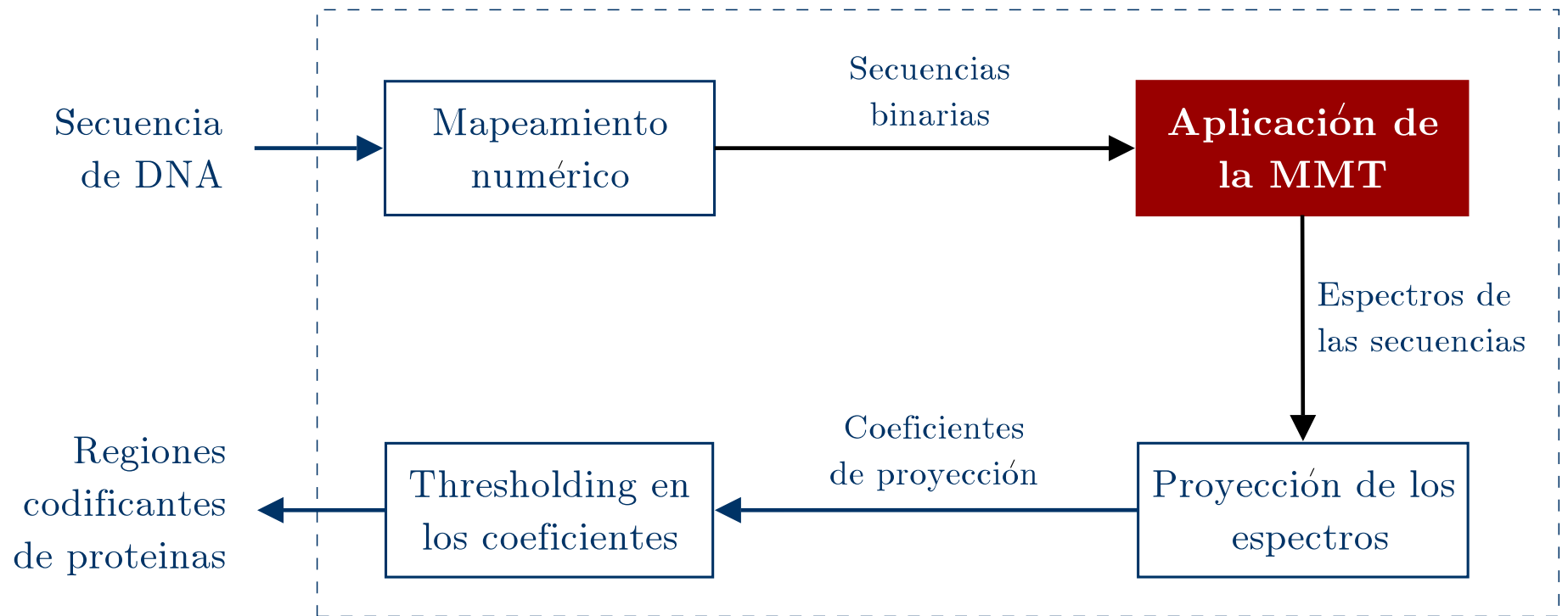
# Mapeamiento Numérico de Nucleótidos

Uso de reglas de mapeamiento para la creación de secuencias binarias a partir de secuencias simbólicas.

Regla	Atribución			
	A	C	G	T
Base A	1	0	0	0
Base C	0	1	0	0
Base G	0	0	1	0
Base T	0	0	0	1

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.

Secuencia $s$	A	T	G	C	T	T	G	A	C	T
$u_A$	1	0	0	0	0	0	0	1	0	0
$u_C$	0	0	0	1	0	0	0	0	1	0
$u_G$	0	0	1	0	0	0	1	0	0	0
$u_T$	0	1	0	0	1	1	0	0	0	1



# Aplicación de la MMT

La MMT con diferentes escalas  $a$  y frecuencia angular  $\omega_0$ , siendo un múltiplo de tres, es calculada para todas las secuencias binarias utilizando  $\psi_{\text{MMT}}$ .

$$U_A(b, a) = \int u_A(x) \psi_{\text{MMT}}(x, b, a) dx$$

$$U_C(b, a) = \int u_C(x) \psi_{\text{MMT}}(x, b, a) dx$$

$$U_G(b, a) = \int u_G(x) \psi_{\text{MMT}}(x, b, a) dx$$

$$U_T(b, a) = \int u_T(x) \psi_{\text{MMT}}(x, b, a) dx$$

Transformadas con diferentes escalas pueden ser aplicadas para el análisis de secuencias de ADN. Los mejores resultados fueron para escalas separadas exponencialmente entre 0,2 y 0,7.

# Espectros de las Secuencias

El espectro de cada secuencia binaria es definido como el módulo al cuadrado de sus coeficientes después de ser aplicada la transformada:

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

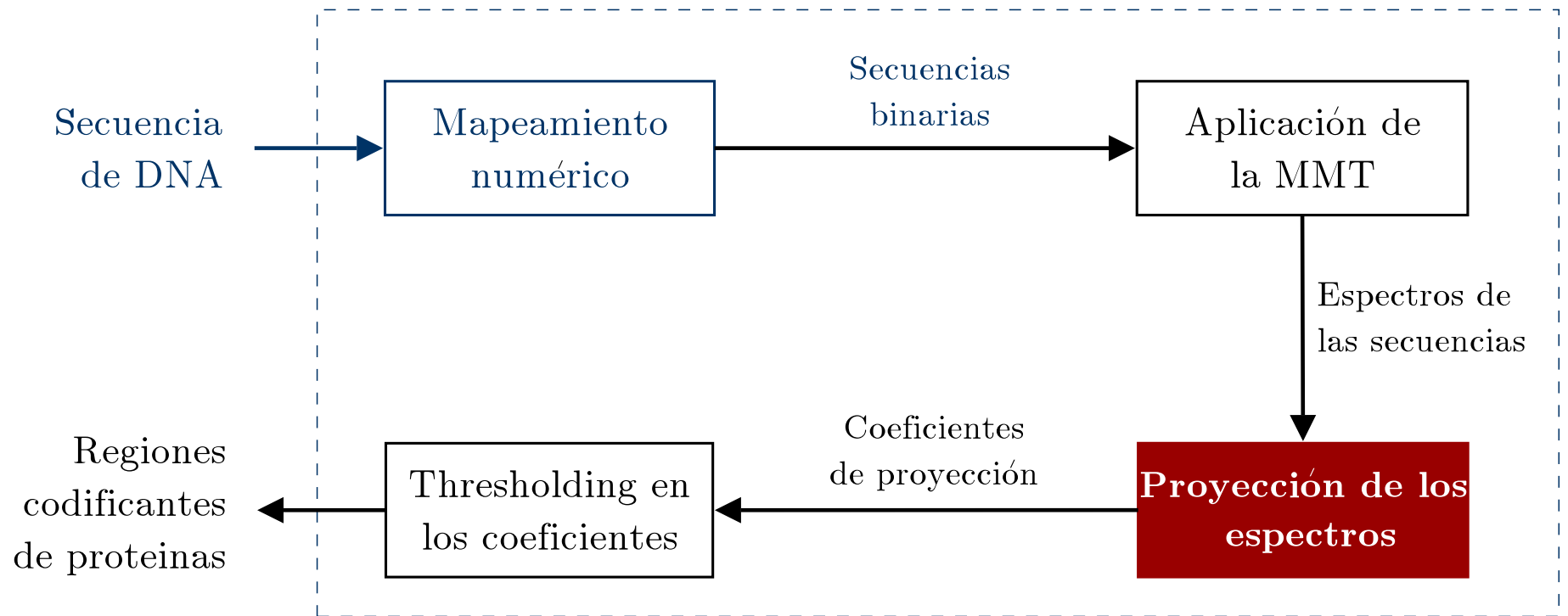
$$m_C(b, a) = |U_C(b, a)|^2$$

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

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

Así, el espectro total, que combina las contribuciones de todas las transformadas, es dada por:

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



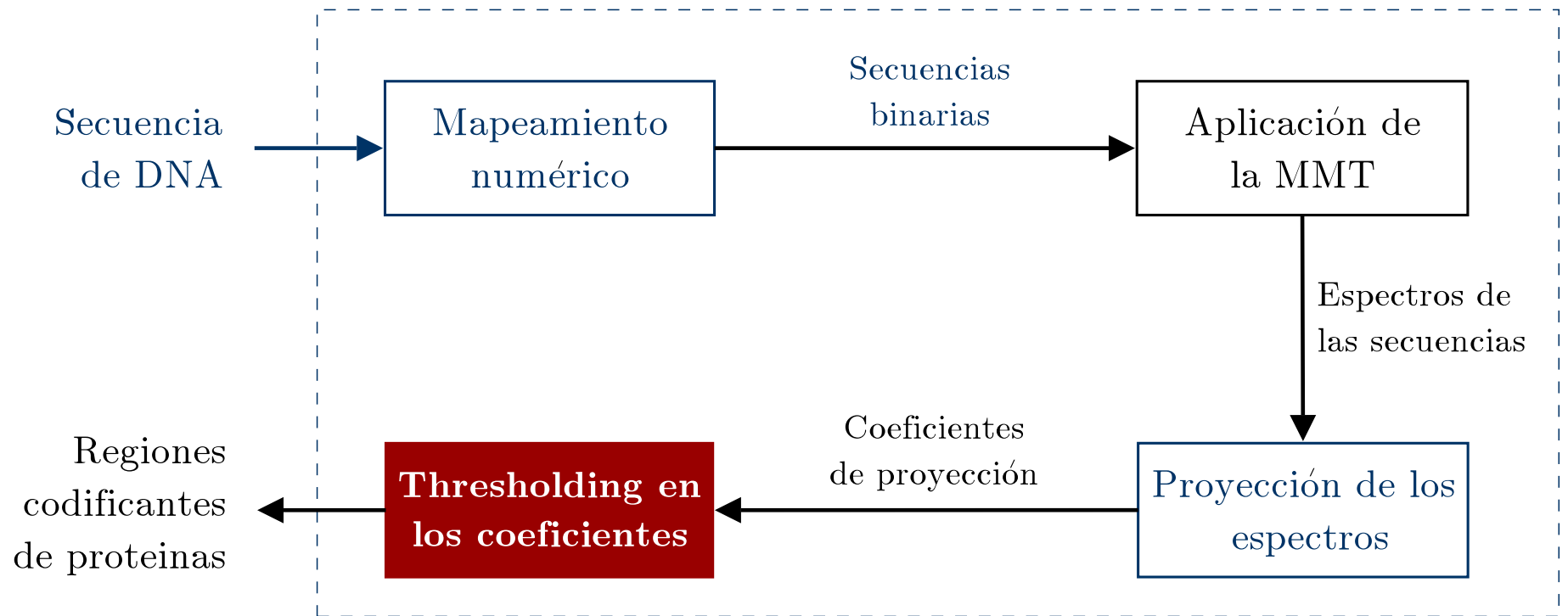
# Proyección de los Espectros

El espectro total de la secuencia analizada es proyectado en el eje de las posiciones. Dada una secuencia de tamaño  $N$ , los coeficientes de proyección del espectro total será dado por:

$$M_p(b) = \sum_a M(b, a), \quad 1 \leq b \leq N$$

La proyección en el eje de las escalas revela cual escala mantiene mas energía en la secuencia através de las posiciones:

$$M_s(a) = \sum_{b=1}^N M(b, a)$$



# ***Thresholding* en los Coeficientes de Proyección**

*Thresholding* sobre  $M_p$  permite excluir posiciones donde los coeficientes sean pequeños, i.e., todos los coeficientes menores que un valor dado son substituidos por cero (*threshold* porcentual).

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 regiones codificantes de la secuencia.

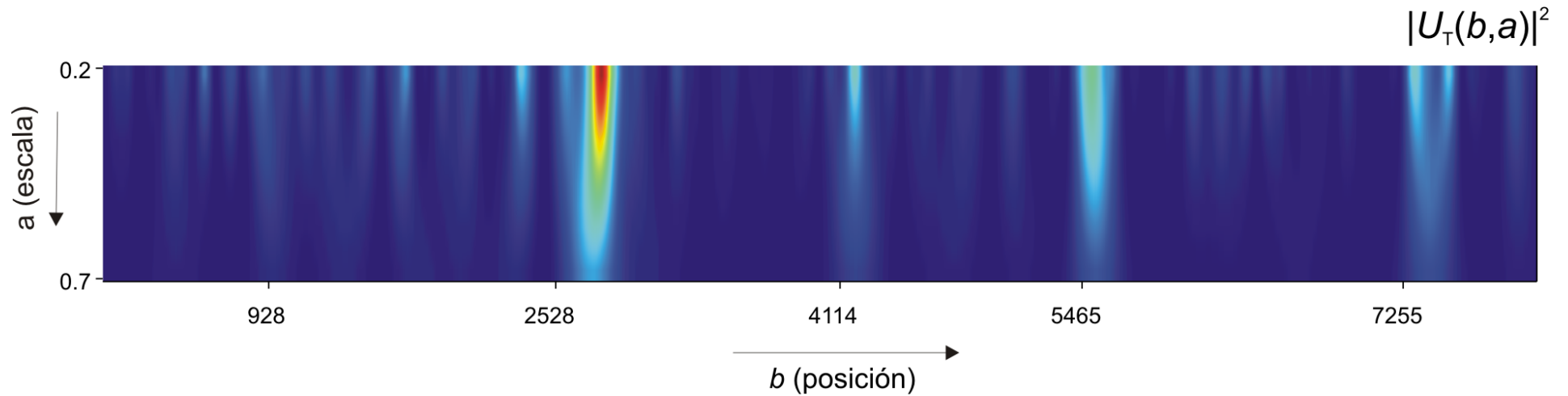
# Conjunto de Secuencias Utilizado

Enfocamos nuestro estudio en el análisis de secuencias de ADN sintéticas y reales.

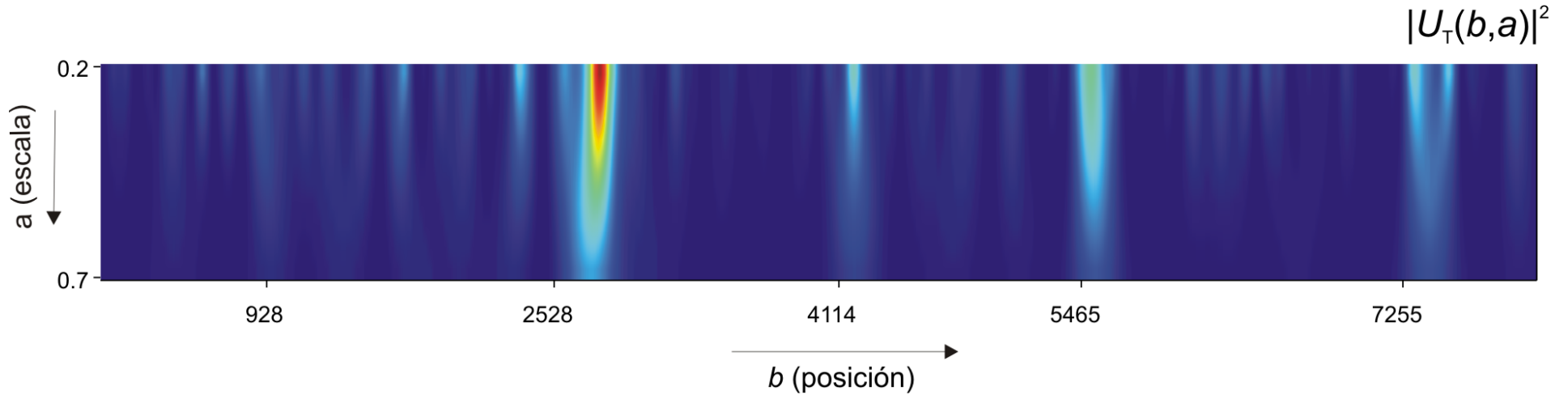
En esta presentación tratamos únicamente a un conjunto de 570 secuencias de vertebrados con sus respectivos límites entre exones e intrones [ **Burset & Guigó, 1996**].

Región	Cantidad	Bases	Tamaño	
			Promedio	Desviación
Éxon	2649	444498 <b>(15.4 %)</b>	<b>168</b>	<b>222</b>
Íntron	2079	1310452 (45.3 %)	630	909
Inter-genica	1132	1137199 (39.3 %)	1004	1464
Total	5860	2892149	-	-

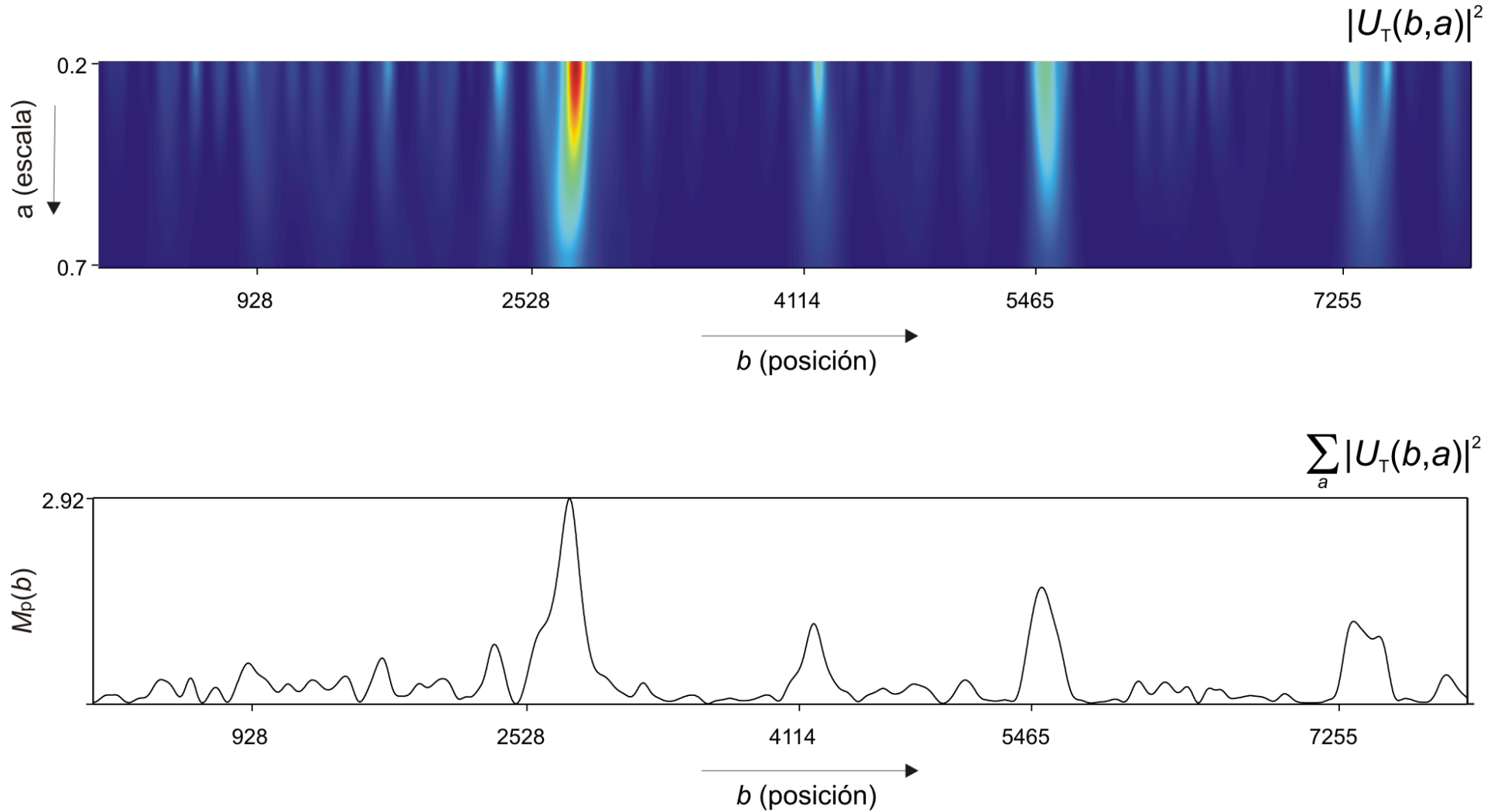
# Resultados: Secuencia F56F11.4



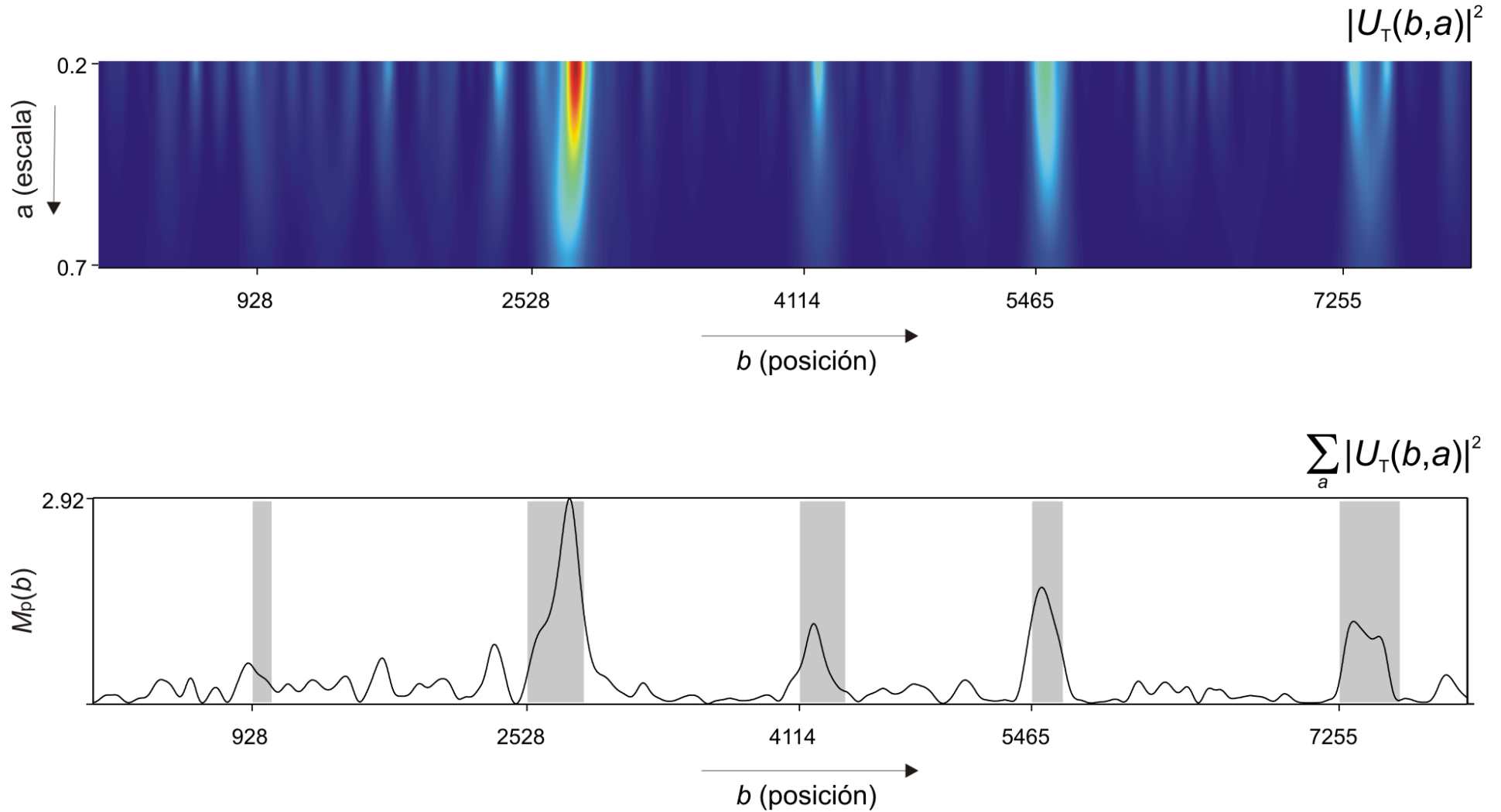
# Resultados: Secuencia F56F11.4



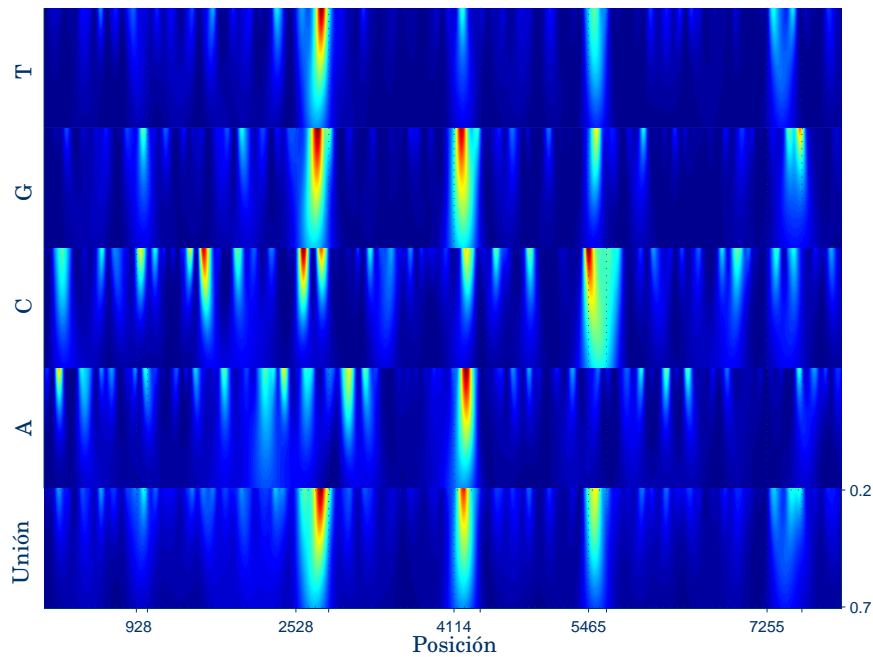
# Resultados: Secuencia F56F11.4



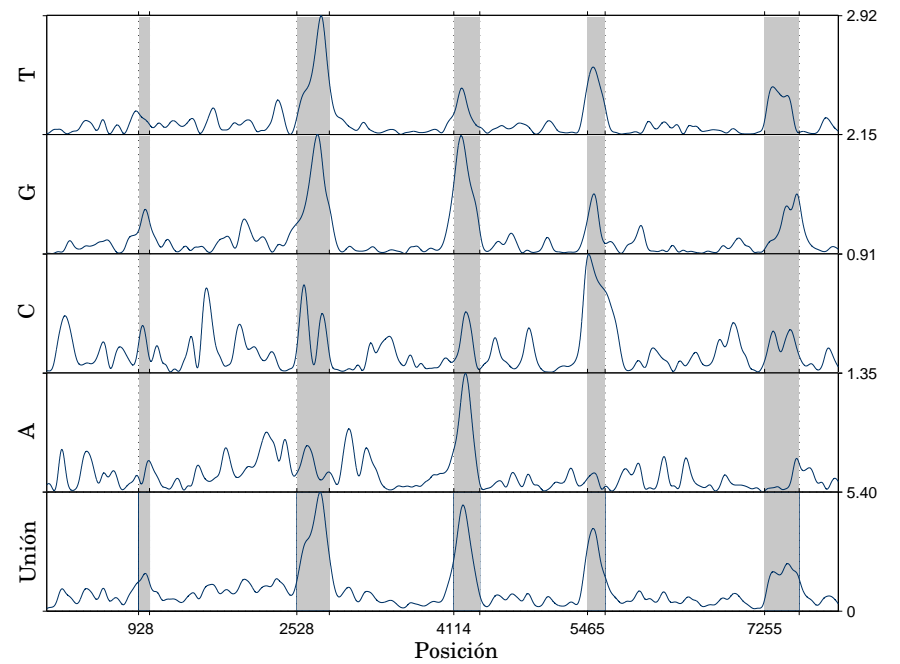
# Resultados: Secuencia F56F11.4



# Resultados: Secuencia F56F11.4

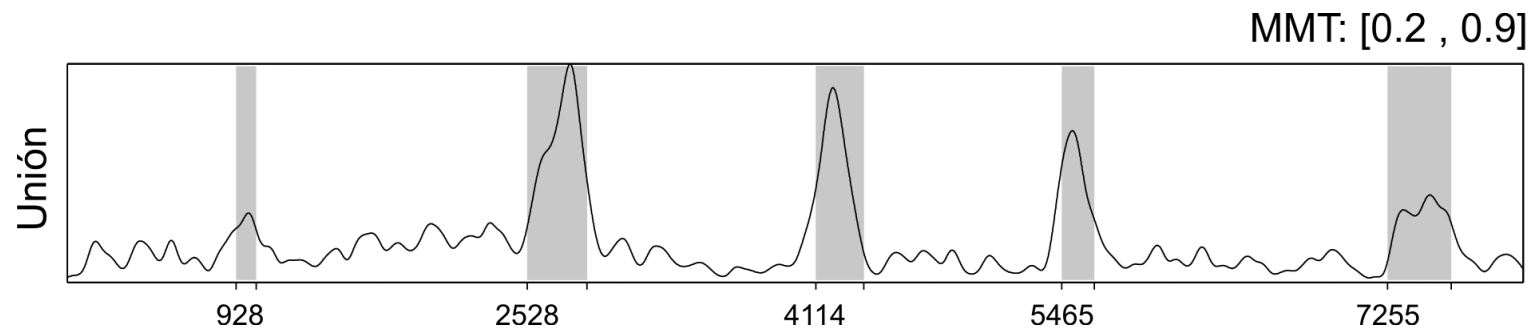


Espectrogramas

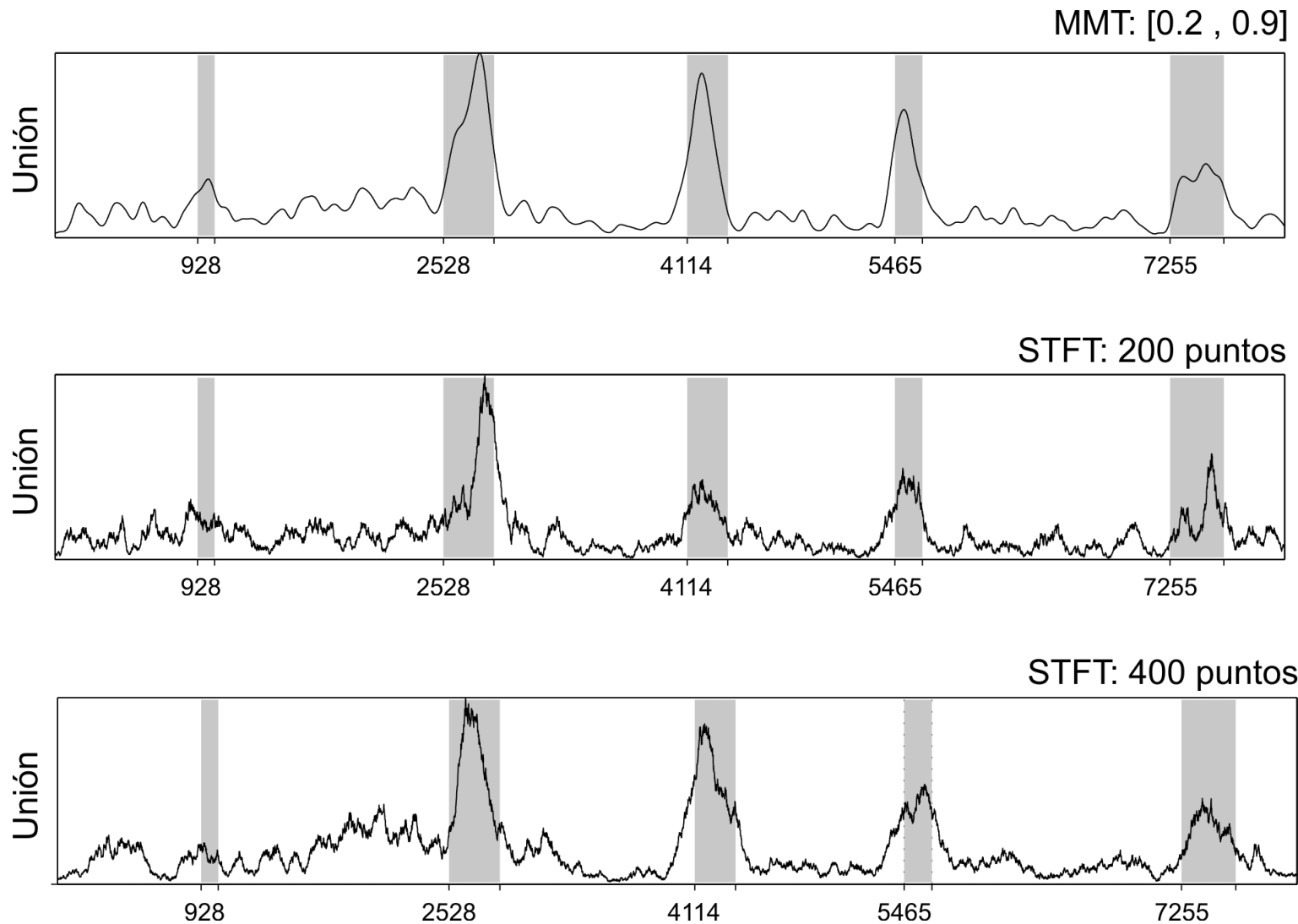


Proyecciones

# Resultados: Secuencia F56F11.4



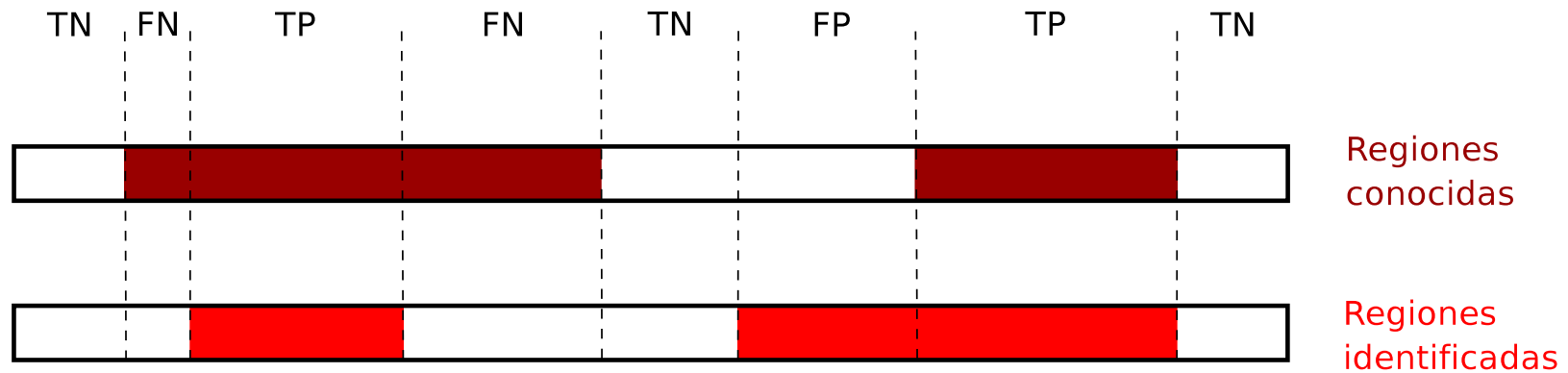
# Resultados: Secuencia F56F11.4



# Medidas de Desempeño

Las medidas de exactitud en el nivel de los nucleótidos, proponen una **forma de comparación** de regiones identificadas con regiones codificantes conocidas.

La medición de regiones identificadas contra regiones codificantes conocidas es realizada mediante conteo de nucleótidos.



# Medidas de Desempeño

- **Sensibilidad** ( $S_n$ ), proporción de nucleótidos codificantes correctamente identificados como codificantes.

$$S_n = \frac{TP}{TP+FN}$$

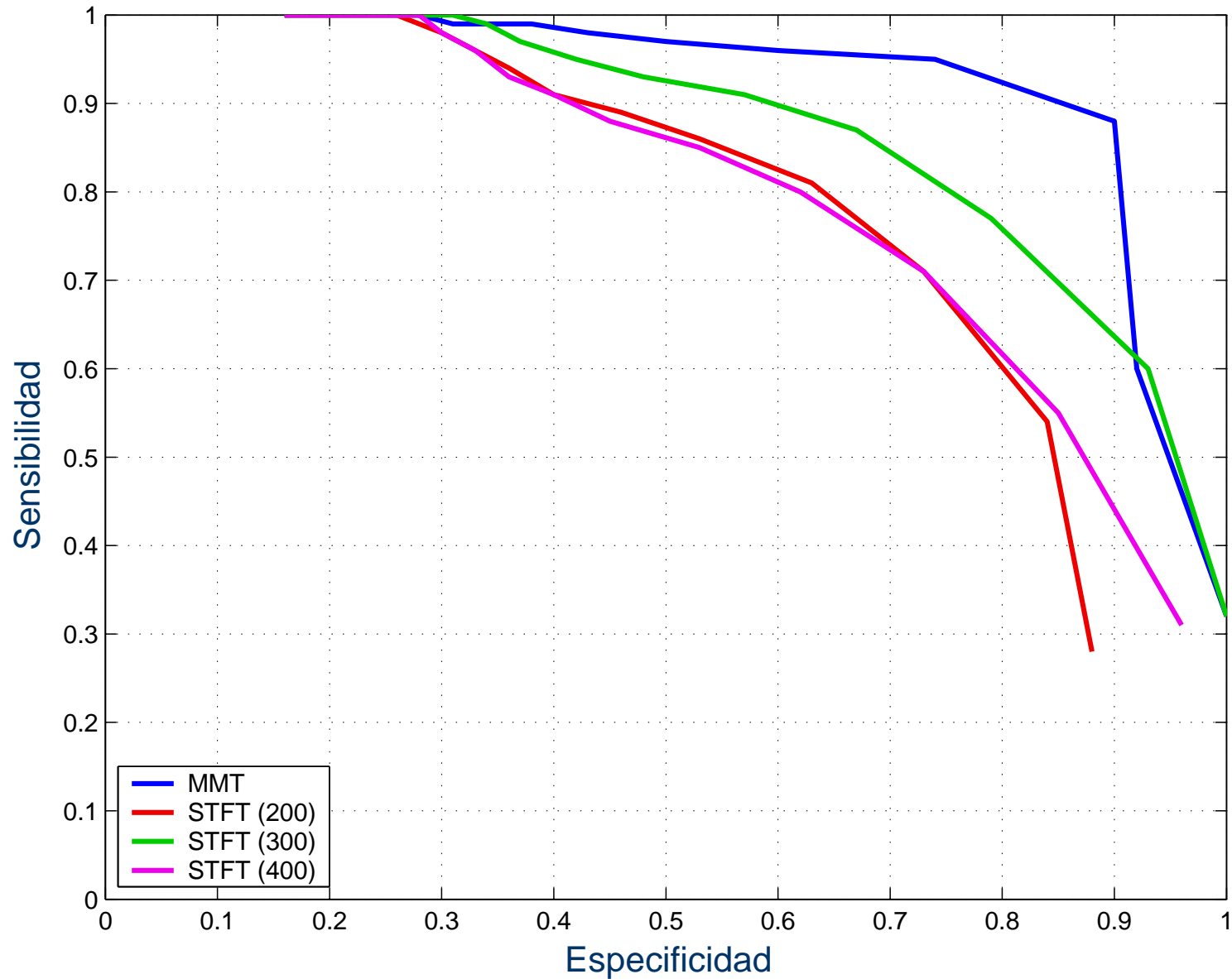
- **Especificidad** ( $S_p$ ), proporción de nucleótidos identificados como codificantes que son actualmente codificantes.

$$S_p = \frac{TP}{TP+FP}$$

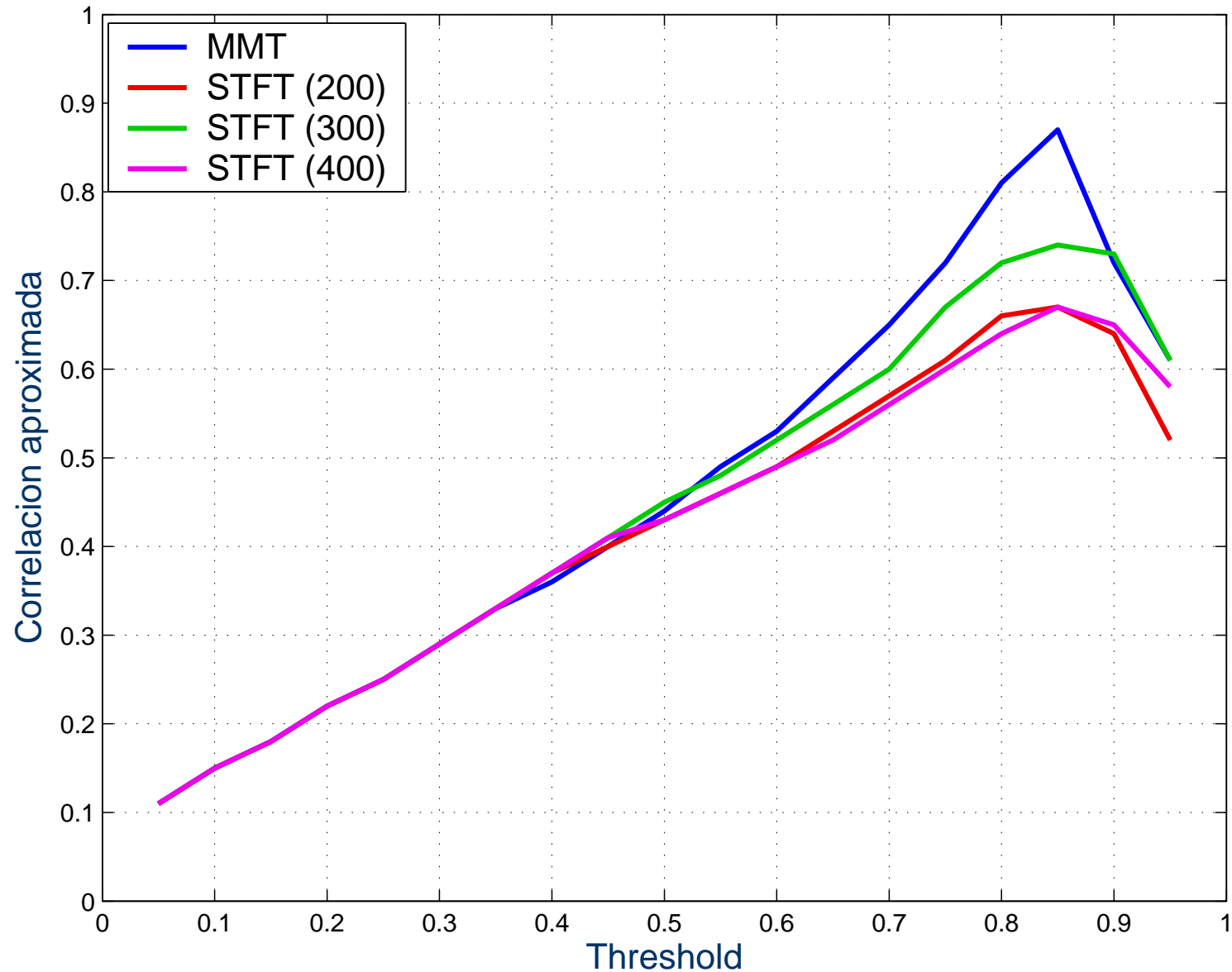
- **Correlación aproximada** ( $AC$ ), medida que combina a  $S_n$  y  $S_p$ .

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

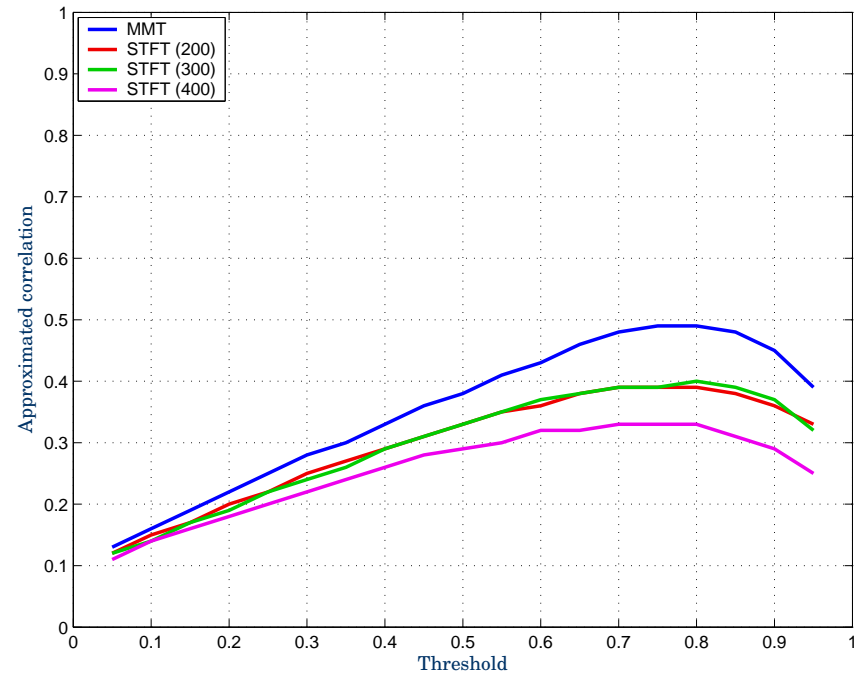
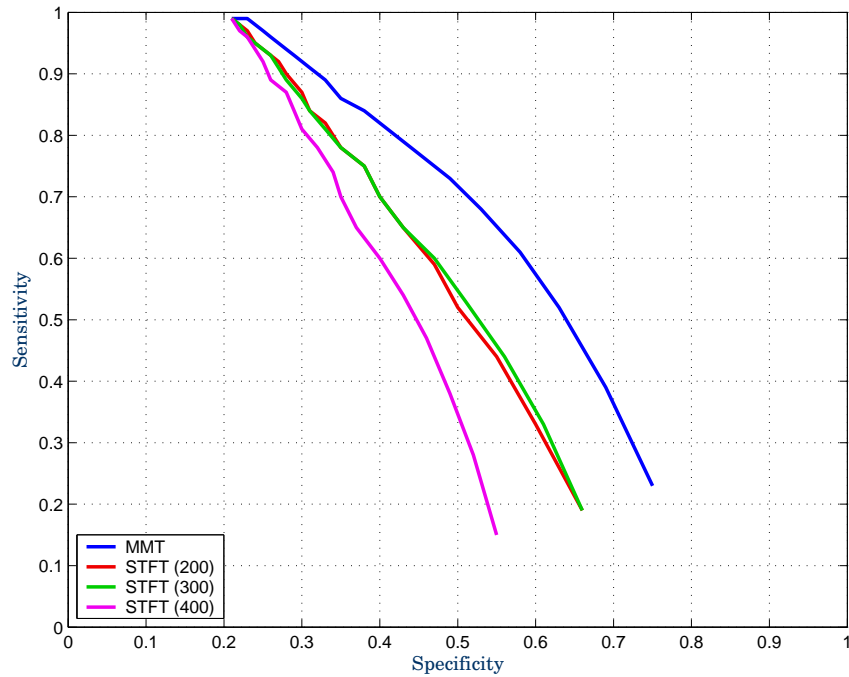
# Desempeño: Secuencia F56F11.4



# Desempeño: Secuencia F56F11.4

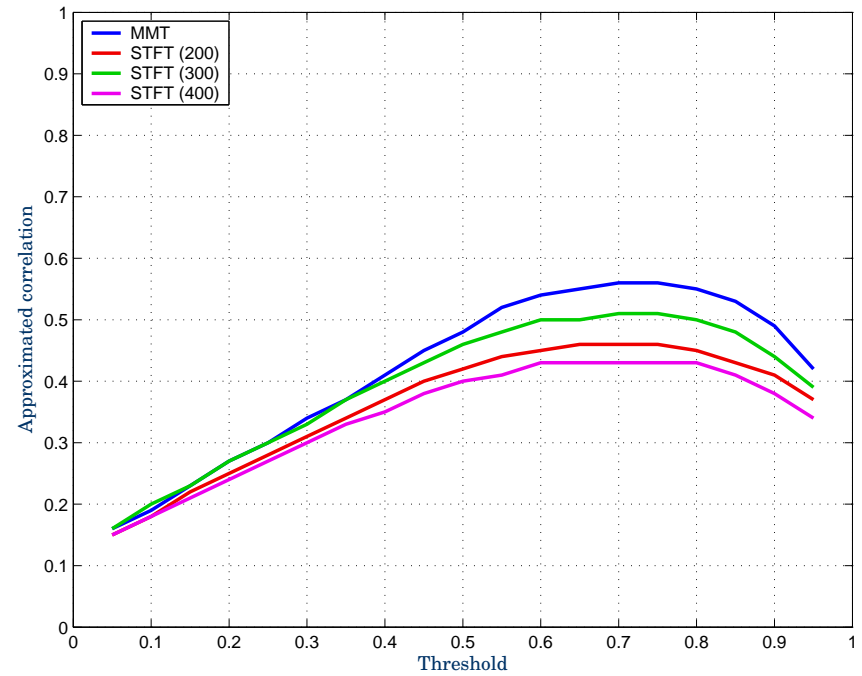
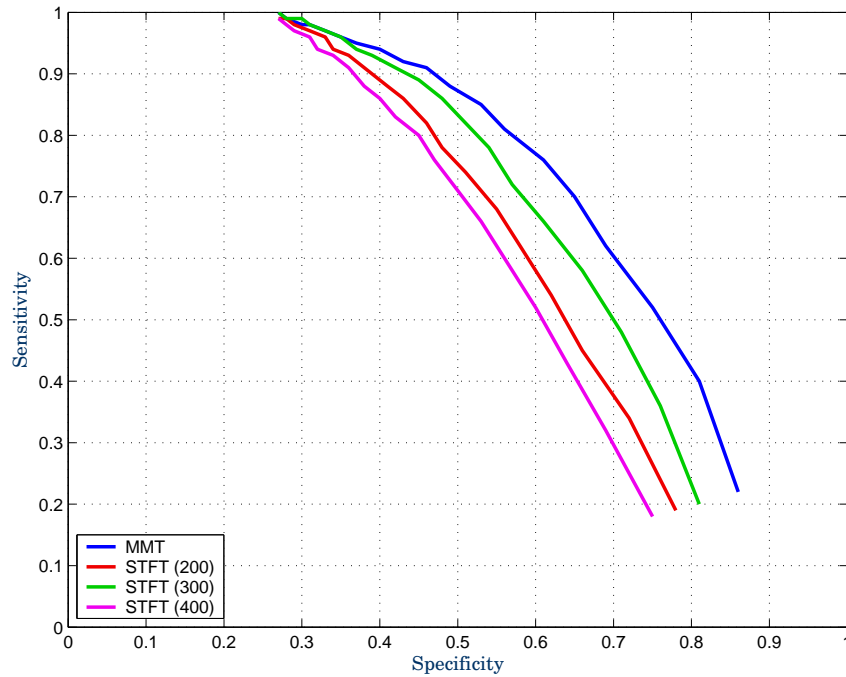


# Desempeño: Conjunto de Secuencias



570 secuencias

# Desempeño: Conjunto de Secuencias



103 secuencias.  
Subconjunto con exones **mayores a 100bp**.

# Software

<http://www.vision.ime.usp.br/~jmena/DSPgenomics/>

**Identification of Protein Coding Regions Through the Modified Morlet Transform**

Query data  
Select the sequence file in [fasta format](#) you wish to transform

Submit Reset

Parameters

First scale:	0.2
Last scale:	0.7
Number of scales:	20
Threshold value:	0.85

Some [DNA sequences](#) for test.  
[back to home](#)

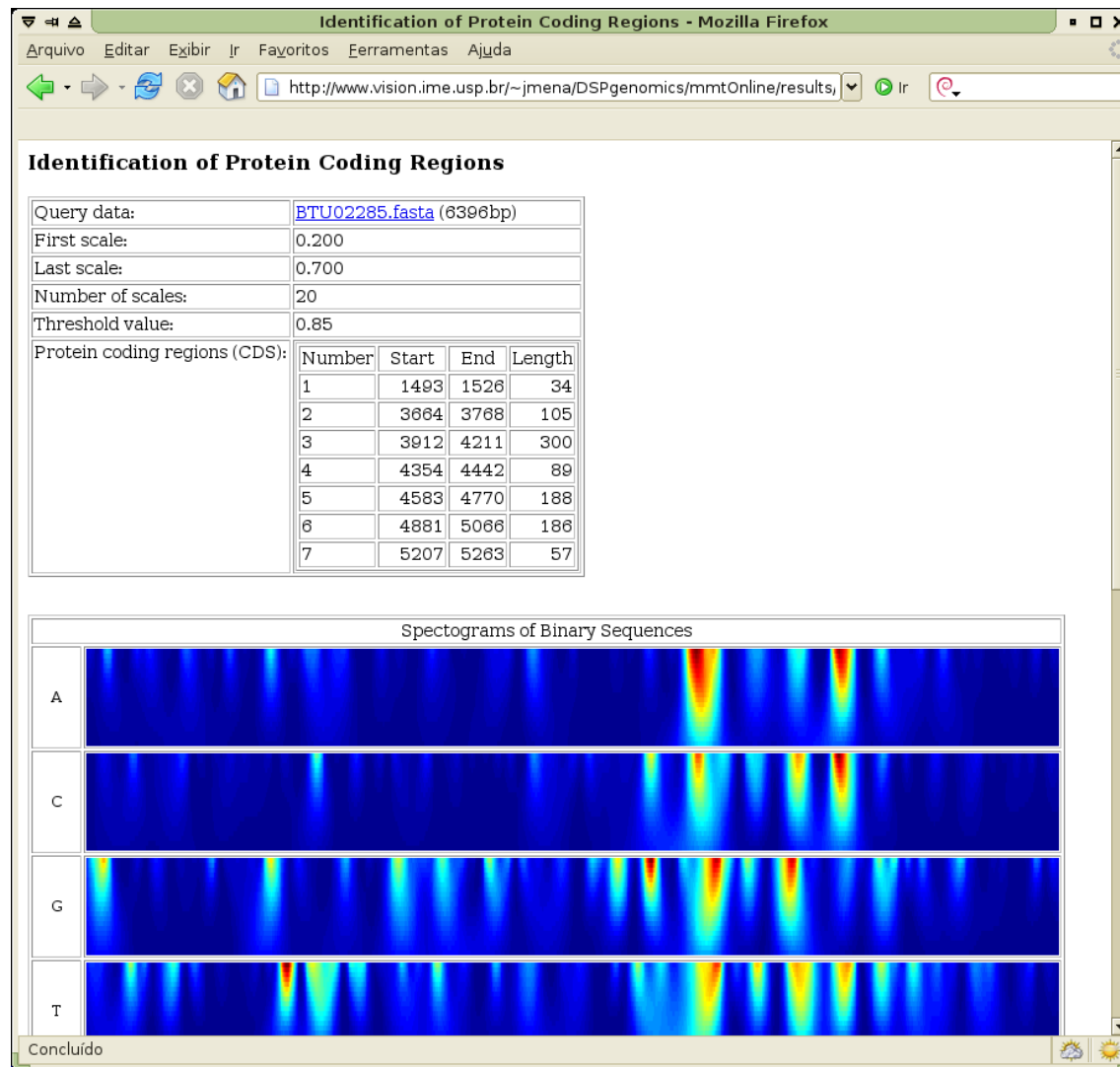
Bugs, suggestions or comments feel free to email me: [jmena AT vision.ime.usp.br](mailto:jmena AT vision.ime.usp.br)

[J. P. Mena-Chalco](#)  
Last modified: Qui Mar 2 14:30:31 BRT 2006

[http://www.vision.ime.usp.br/~jmena/DSPgenomics/mmtOnline/fasta\\_format.html](http://www.vision.ime.usp.br/~jmena/DSPgenomics/mmtOnline/fasta_format.html)

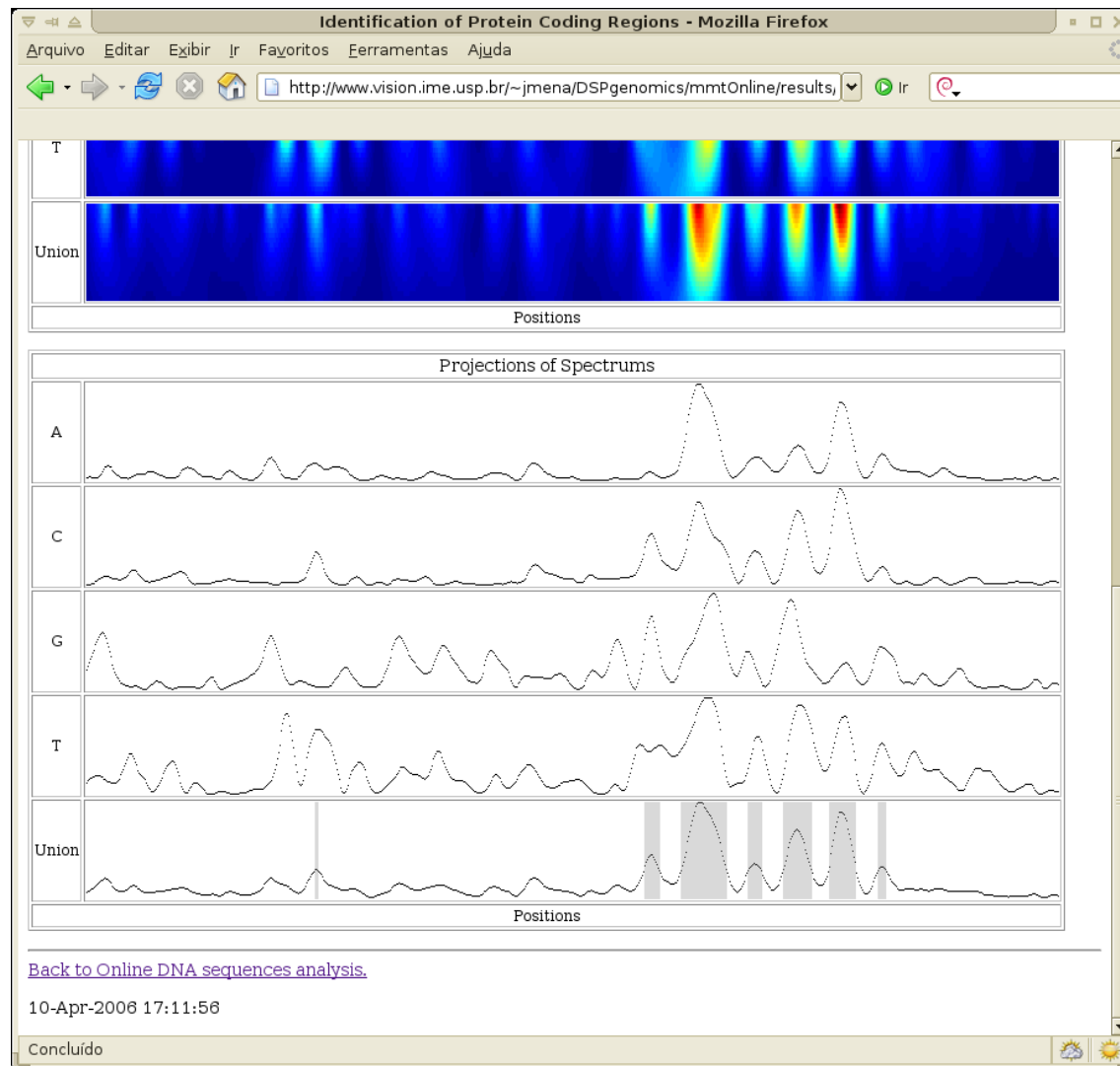
# Software

<http://www.vision.ime.usp.br/~jmena/DSPgenomics/>



# Software

<http://www.vision.ime.usp.br/~jmena/DSPgenomics/>



# Discusión

- Mejores desempeños obtenidos con un *threshold* de 85 %.

# Discusión

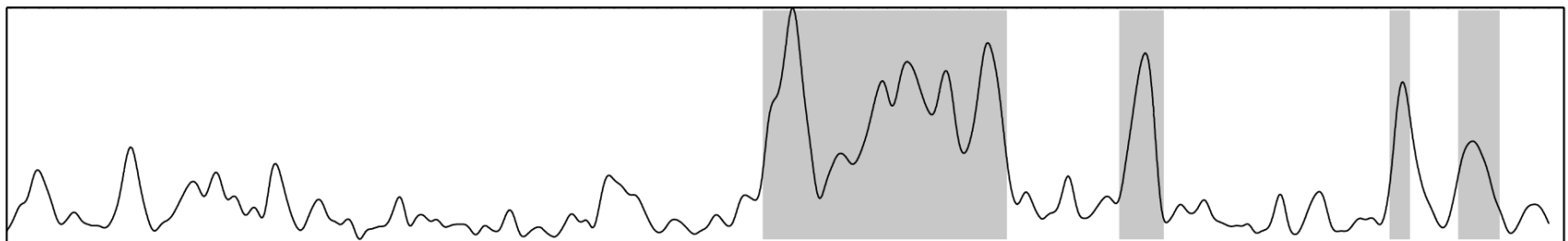
- Mejores desempeños obtenidos con un *threshold* de 85 %.
- Nivel máximo de exactitud alcanzado de 56 %.

# Discusión

- Mejores desempeños obtenidos con un *threshold* de 85 %.
- Nivel máximo de exactitud alcanzado de 56 %:
  1. Existencia de TBP no uniforme en las regiones codificantes.

# Discusión

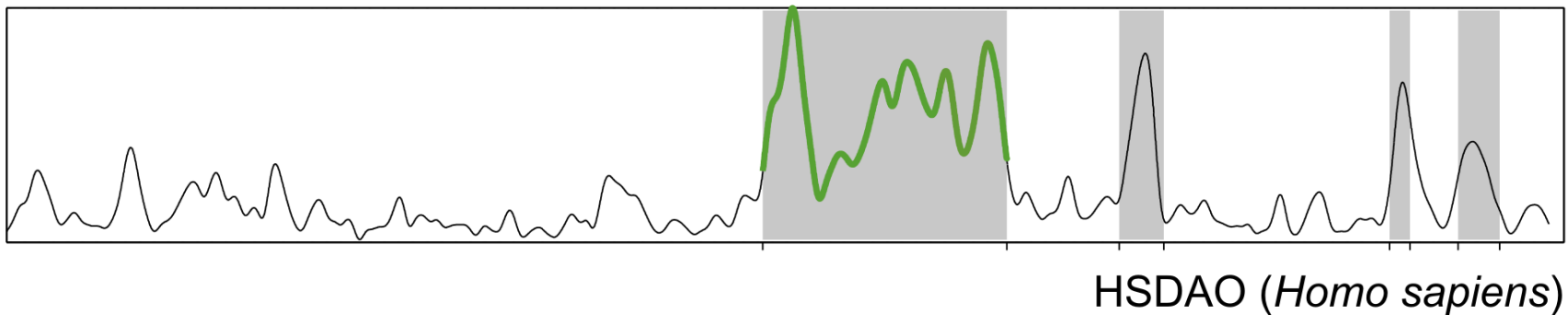
- Mejores desempeños obtenidos con un *threshold* de 85 %.
- Nivel máximo de exactitud alcanzado de 56 %:
  1. Existencia de TBP no uniforme en las regiones codificantes.



HSDAO (*Homo sapiens*)

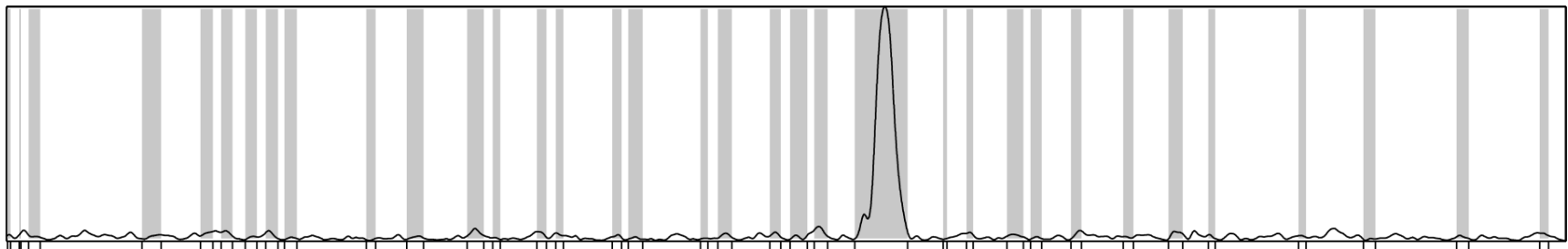
# Discusión

- Mejores desempeños obtenidos con un *threshold* de 85 %.
- Nivel máximo de exactitud alcanzado de 56 %:
  1. Existencia de TBP no uniforme en las regiones codificantes.



# Discusión

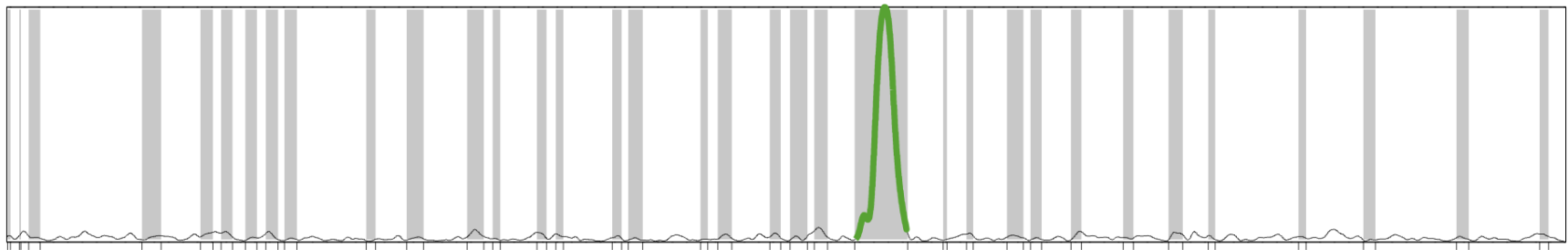
- Mejores desempeños obtenidos con un *threshold* de 85 %.
- Nivel máximo de exactitud alcanzado de 56 %:
  1. Existencia de TBP no uniforme en las regiones codificantes.
  2. Existencia de poca o ninguna TBP en las regiones codificantes.



GGVITIIG (*Gallus gallus*)

# Discusión

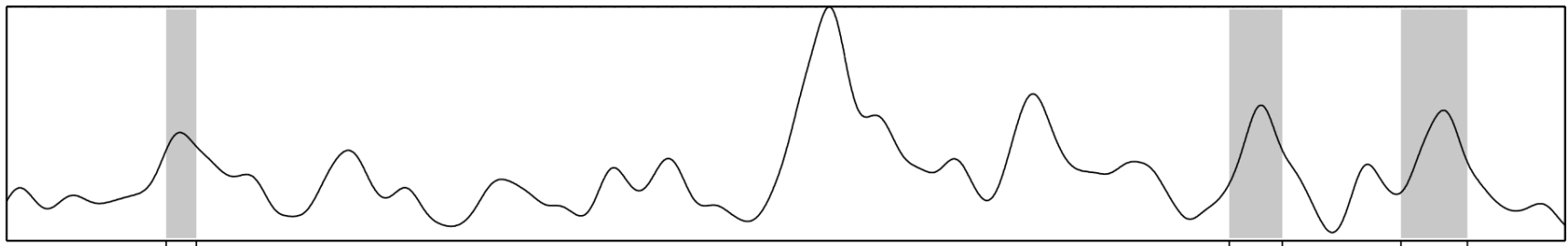
- Mejores desempeños obtenidos con un *threshold* de 85 %.
- Nivel máximo de exactitud alcanzado de 56 %:
  1. Existencia de TBP no uniforme en las regiones codificantes.
  2. Existencia de poca o ninguna TBP en las regiones codificantes.



GGVITIIG (*Gallus gallus*)

# Discusión

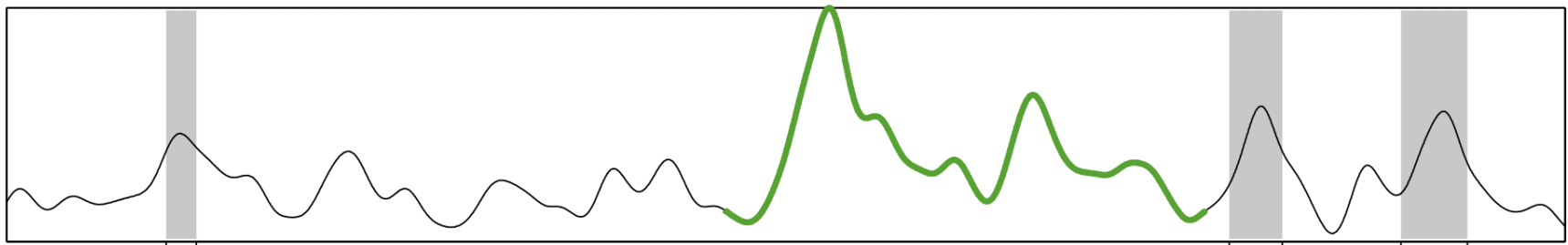
- Mejores desempeños obtenidos con un *threshold* de 85 %.
- Nivel máximo de exactitud alcanzado de 56 %:
  1. Existencia de TBP no uniforme en las regiones codificantes.
  2. Existencia de poca o ninguna TBP en las regiones codificantes.
  3. Existencia de TBP en las regiones no codificantes.



MMACLGNA (*Mus musculus*)

# Discusión

- Mejores desempeños obtenidos con un *threshold* de 85 %.
- Nivel máximo de exactitud alcanzado de 56 %:
  1. Existencia de TBP no uniforme en las regiones codificantes.
  2. Existencia de poca o ninguna TBP en las regiones codificantes.
  3. Existencia de TBP en las regiones no codificantes.



MMACLGNA (*Mus musculus*)

# Conclusiones

- El método se basa únicamente en la TBP existente en las regiones codificantes. No es usada ninguna otra información adicional.
- El análisis comparativo usando secuencias reales de ADN muestra que el método propuesto usando la MMT tiene un desempeño superior sobre los otros basados en la STFT.
- El método es flexible y robusto a variaciones de escala para el análisis de secuencias de ADN.
- El método brinda una forma gráfica de representación de la TBP encontrada localmente en las regiones codificantes.

# Agradecimientos

- A la profesora Helaine Carrer por la orientación relativa a los aspectos biológicos.
- A los profesores Yossi Zana y Luiz Velho por las discusiones y sugerencias en este trabajo.
- A la CAPES por el apoyo financiero en el Programa de Estudante Convênio (PEC-PG), Perú-Brasil.