GENERALIZACION DE LA DIVERGENCIA DE JENSEN SHANNON A ESTADISTICA NO EXTENSIVA PARA EL ANALISIS DE SECUENCIAS

NONEXTENSIVE GENERALIZATION OF THE JENSEN SHANNON DIVERGENCE FOR SEQUENCE ANNALYSIS

Diego Bussandri[1], Leonel Garro Linck[1], Miguel Ré[1,2] y Pedro Lamberti[1,3]

[1]Facultad de Matemática, Astronomía y Física, Universidad Nacional de Córdoba
Ciudad Universitaria - (5010) - Córdoba - Argentina
[2]Facultad Regional Córdoba, Universidad Tecnológica Nacional.
Maestro López y Cruz Roja Argentina - (5010) - Córdoba - Argentina
[3]CONICET
e-mail: mre@cbasicas.frc.utn.edu.ar, re@famaf.unc.edu.ar

Recibido: 17/12/12 – Aceptado: 16/09/13

Resumen

La divergencia de Jensen Shannon (JSD), una versión simetrizada de la divergencia de Kullback Leibler, permite cuantificar la diferencia entre distribuciones de probabilidad. Debido a esta propiedad ha sido ampliamente utilizada para el análisis de secuencias simbólicas, comparando la composición simbólica de posibles subsecuencias. Una ventaja que ofrece JSD es que no requiere el mapeo de la secuencia simbólica a una secuencia numérica, necesaria por ejemplo en el análisis de correlación espectral.

Se han propuesto distintas extensiones de JSD para mejorar la detección de bordes de subsecuencias en una secuencia, en particular para el análisis de secuencias de DNA.

Desde su propuesta original, la extensión propuesta por Tsallis a la entropía de Boltzmann Gibbs ha sido considerada para extender sus resultados y aplicaciones. Sin embargo no surge una única posibilidad para la extensión de JSD a partir de la definición de Tsallis.

Consideramos aquí posibles extensiones de la JSD en el marco de la entropía de Tsallis y consideramos los resultados que se obtienen cuando se aplican al análisis de secuencias simbólicas para la detección de bordes de subsecuencias.

Palabras Clave: Segmentación, distancias entrópicas, Jensen-Shannon, entropía no extensiva

Abstract

Jensen Shannon Divergence (JSD), a symmetrized version of the Kullback Leibler divergence, allows to quantify the difference between probability distributions. Due to this property, JSD has been widely used in the symbolic sequence annalysis by comparing the symbolic composition of possible subsequences. One advantage of JSD is that it does not require the symbolic sequence to be mapped to a numerical sequence, which is necessary for instance in spectral correlation analysis.

Different generalizations of JSD have been proposed to improve detection of sequences borders, in particular for DNA sequence analysis.

Since its original proposal, Tsallis entropy has been considered to generalize Boltzmann Gibbs Shannon entropy results and applications. Different JSD Tsallis extensions have been suggested and its properties analyzed.

We present here possible extensions of JSD in Tsallis entropy framework and consider the results obtained when applied to DNA sequence analysis for subsequences border detection.

Keywords: Segmentation, entropic distances, Jensen-Shannon, non extensive entropy

1 Introducción

Diversos problemas en Mecánica Cuántica, Mecánica Estadística, Teoría de la Información o Biofísica pueden plantearse en forma conveniente en términos de una *divergencia*: una regla para establecer una comparación cuantitativa entre dos o más distribuciones de probabilidad. Siguiendo distintos argumentos, se han introducido medidas de divergencia alternativas en la literatura, entre las que mencionamos la divergencia de Kullback-Leibler, la de Jeffreys o la de Jensen-Shannon; o las distancias de Wooters, o Bures. Estas funcionales han sido utilizadas para el estudio de series temporales, establecer medidas de complejidad o

distinguir entre estados cuánticos.
En particular el análisis estadístico de secuencias simbólicas es importante en diversos campos como lingüística o análisis de secuencias de ADN. El uso de funcionales provenientes de la teoría de la información para el análisis de secuencias simbólicas presenta la ventaja de que no es necesario el mapeo a una secuencia numérica como en el análisis espectral o de correlación.
La divergencia de Jensen-Shannon (JSD) es una de estas funcionales que permite cuantificar la diferencia entre dos (o inclusive más) distribuciones de probabilidad. JSD presenta además la ventaja adicional de no requerir la continuidad absoluta de las distribuciones a comparar. Así la JSD puede utilizarse para comparar la composición de los símbolos de dos secuencias en principio diferentes asociando la frecuencia de aparición de cada símbolo con distribuciones de probabilidad.
En su propuesta de generalización de la estadística de Boltzamnn-Gibbs, Tsallis[1] incorpora una definición no extensiva de entropía a través de una deformación de la función logaritmo

$$\mathrm{lq}\,(x) = \frac{x^{1-q} - 1}{1-q} \tag{1}$$

que se reduce a la función $\ln(x)$ en el límite $q \to 1$. Desde la formulación de esta propuesta, ha despertado gran interés la generalización de resultados y aplicaciones basados en entropía. En particular distintas generalizacones de la JSD han sido consideradas y sus propiedades analizadas. Martins *et al.*[2] han presentado algunas de estas posibles generalizaciones. A su vez, en su trabajo sobre análisis de secuencias simbólicas, Grosse *et al.*[3] dan tres interpretaciones intuitivas de la JSD que permiten caracterizar las generalizaciones propuestas para la JSD.
En esta comunicación presentamos un estudio de la aplicación de estas generalizaciones al análisis de secuencias simbólicas. En particular se considera la detección en una secuencia de la posición de bordes de subsecuencias de distinta composición.
En la próxima sección se presenta la definición de la JSD y se incluyen las interpretaciones intuitivas antes mencionadas. En la sección 3 se presentan las posibles generalizaciones de la JSD a la estadística de Tsallis[1], siguiendo las interpretaciones intuitivas. Estas interpretaciones pueden ser de utilidad al considerar nuevas generalizaciones de la JSD como en el caso de modelos markovianos[4] a fin de detectar la correlación en la aparición de los símbolos en una dada secuencia. En la sección 4 se presentan el método de segmentación y los resultados obtenidos en experimentos de control con secuencias generadas a partir de distribuciones conocidas. Finalmente en la sección 5 presentamos un resumen y las conclusiones de esta comunicación.

2 Divergencia de Jensen-Shannon

Diversas medidas han sido propuestas para cuantificar la diferencia (o divergencia) entre dos distribuciones de probabilidad. Consideramos aquí la divergencia de Jensen-Shannon definida como sigue: denotamos por $\mathbf{p}_i$ la distribución de probabilidad para un conjunto de k símbolos, de manera que $p_i(e_k)$ es la probabilidad del símbolo e_k en la distribución i y por $\pi(i)$ el peso de la distribución i; que deben satisfacer las restricciones

$$\sum_{j=1}^{k} p_i(e_k) = 1 \qquad \sum_{i=1}^{2} \pi(i) = 1 \tag{2}$$

Se define la divergencia de Jensen-Shannon (JSD) entre las distribuciones $\mathbf{p}_i$ con sus respectivos pesos por

$$D[\mathbf{p}_1, \mathbf{p}_2] \equiv H\left[\sum_{i=1}^{2} \pi(i)\,\mathbf{p}_i +\right] - \sum_{i=1}^{2} \pi(i)\,H[\mathbf{p}_i] \tag{3}$$

siendo

$$H[\mathbf{p}_i] = -\sum_{j=1}^{k} p_i(e_j) \ln(p_i(e_j)) \tag{4}$$

la entropía de Boltzmann-Gibbs-Shannon (BGSe) de la distribución $\mathbf{p}_i$.
La JSD satisface las siguientes propiedades:

1. partiendo de la desigualdad de Jensen para funciones cóncavas

$$D[\mathbf{p}_1, \mathbf{p}_2] \geq 0 \tag{5}$$

con la igualdad válida si y sólo si $\mathbf{p}_1 = \mathbf{p}_2$

2. D es simétrica en sus argumentos

$$D[\mathbf{p}_1, \mathbf{p}_2] = D[\mathbf{p}_2, \mathbf{p}_1] \tag{6}$$

3. D está bien definida aún cuando $\mathbf{p}_i$ no sean absolutamente continuas.

Cabe señalar además que D puede extenderse a más de dos distribuciones manteniendo las propiedades enunciadas.
A la JSD pueden darse tres interpretaciones intuitivas[3]:

2.1 Interpretación desde la física estadística

D puede interpretarse como la entropía de mezcla a partir de la siguiente consideración: supongamos 2 recipientes conteniendo una mezcla de k gases ideales y sea $\mathbf{f}_i$ el vector de las fracciones molares de los gases en el recipiente i. Si el número total de moléculas en el recipiente i es n_i y $N = n_1 + n_2$ entonces la entropía de mezcla al unir los recipientes es

$$\begin{aligned} S_m &= k_B\left[NH[\mathbf{f}] - \textstyle\sum_{i=1}^{2} n_i H[\mathbf{f}_i]\right] \\ &= Nk_B D[\mathbf{f}_1, \mathbf{f}_2] \end{aligned} \tag{7}$$

siendo $\mathbf{f} = \sum_{i=1}^{2} (n_i/N)\,\mathbf{f}_i$ y k_B la constante de Boltzmann.
Los pesos se eligen como $\pi(i) = n_i/N$ en esta interpretación .

2.2 Interpretación desde la teoría de información

D puede interpretarse como la información mutua: consideremos una secuencia $\mathcal{S}$ conformada por N símbolos extraídos de un alfabeto $\mathcal{A}$ de k elementos e_k. Sea $p(e_k)$ la probabilidad de encontrar el símbolo e_k en una posición dada de la secuencia $\mathcal{S}$. Supongamos que la secuencia $\mathcal{S}$ está dividida en 2 subsecuencias $\mathcal{S}_1, \mathcal{S}_2$ de longitudes n_1 y n_2 respectivamente, con $N = n_1 + n_2$. Supongamos que la secuencia $\mathcal{S}$ no es estacionaria, de manera tal que la probabilidad de encontrar al símbolo e_k en una dada posición de la secuencia $\mathcal{S}_i$ es $p_i(e_k)$. Consideremos ahora las variables aleatorias e con valores en $\mathcal{A}$ y s con valores en $\{\mathcal{S}_1, \mathcal{S}_2\}$. Sea $\hat{p}(e_k, j)$ la probabilidad conjunta de que al elegir el símbolo en la posición n de la secuencia $\mathcal{S}$, este símbolo sea e_k en la subsecuencia $\mathcal{S}_j$. La variable e asume el valor e_k con probabilidad $p(e_k)$ y la variable s el valor j con probabilidad $\pi(j) = n_j/N$. Estas últimas son las probabilidades marginales definidas por

$$p(e_k) \equiv \sum_{j=1}^{2} \hat{p}(e_k, j) \qquad \pi(j) = \sum_{i=1}^{k} \hat{p}(e_i, j) \tag{8}$$

Supongamos la extracción de un elemento e de la secuencia $\mathcal{S}$, pero desconociendo a cuál subsecuencia pertenece. La información mutua en e acerca de la subsecuencia s de pertenencia se define por[5]

$$I(i,j) \equiv -\sum_{i=1}^{k}\sum_{j=1}^{2} \hat{p}(e_i, j) \ln\left(\frac{p(e_i)\,\pi(j)}{\hat{p}(e_i, j)}\right) \tag{9}$$

y da una medida de la ganancia en información de la subsecuencia s a partir del conocimiento del símbolo e. La información mutua es la distancia de Kulback-Leibler entre la probabilidad conjunta y el producto de las probabilidades marginales. Nótese que si y sólo si las variables e y s son estadísticamente independientes, $I(i,j) = 0$.
Tomando en cuenta las definiciones de las probabilidades marginales en (8) y la definición de la probabilidad condicional

$$p_j(e_k) = \frac{\hat{p}(e_k, j)}{\pi(j)} \tag{10}$$

podemos reescribir

$$I(i,j) \equiv -\sum_{i=1}^{k}\sum_{j=1}^{2} \pi(j)\, p_j(e_i) \ln\left(\frac{p(e_i)}{p_j(e_i)}\right) \tag{11}$$

que operando algebraicamente podemos identificar como

$$I(i,j) = D[\mathbf{p}_1, \mathbf{p}_2] \tag{12}$$

siendo $\mathbf{p}_1$ y $\mathbf{p}_2$ las distribuciones de probabilidad para los smbolos en las respectivas subsecuencias, como definidas en (10).

2.3 Interpretación desde la matemática estadística

D puede interpretarse como el cociente del logaritmo de verosimilitud: considérese una secuencia $\mathcal{S}$ conformada por N símbolos donde la probabilidad para el símbolo e_i en una dada posición es $p(e_i)$. Sea F_i el número de veces que se repite e_i en la secuencia. El principio de máxima verosimilitud sugiere que la distribución de probabilidades que hace máxima la verosimilitud

$$L(\mathcal{S} \mid \mathbf{p}) \equiv \prod_{i=1}^{k} p(e_i)^{F_i} \tag{13}$$

es

$$p(e_i) = \frac{F_i}{N} \tag{14}$$

de manera tal que el logaritmo de verosimilitud

$$\ln L = \sum_{i=1}^{k} F_i \ln p_i$$

es máximo con $p_i = f_i = F_i/N$, la frecuencia relativa del símbolo e_i.
Supongamos ahora que la secuencia $\mathcal{S}$ no es estacionaria, sino que está conformada por dos subsecuencias estacionarias $\mathcal{S}_1$ y $\mathcal{S}_2$ de longitudes n_1 y n_2 respectivamente. La máxima verosimilitud para $\mathcal{S}$ en este caso está dada por el producto de verosimilitudes máximas en cada subsecuencia

$$L(\mathcal{S}_j \mid \mathbf{p}_j) \equiv \prod_{i=1}^{k} p_j(e_i)^{F_{i,j}} \tag{15}$$

donde ahora tenemos en principio una distribución distinta para los símbolos e_i en cada subsecuencia, $p_j(e_i)$, y donde $F_{i,j}$ es el número de veces que se repite el símbolo e_i en la subsecuencia $\mathcal{S}_j$. De esta forma obtenemos que la máxima verosimilitud para cada subsecuencia corresponde a

$$p_j(e_i) = \frac{F_{i,j}}{n_j} \tag{16}$$

La diferencia

$$\Delta \ln L_{\max} \equiv \sum_{j=1}^{2} \ln L_{j,\max} - \ln L_{\max} = ND \tag{17}$$

es no negativa y se denomina cociente de verosimilitud logarítmica.

Las interpretaciones presentadas pueden extenderse a más de dos subsecuencias de manera directa.

3 Generalización a estadística no extensiva

Tsallis[1] propone una generalización de la definición de entropía sustituyendo en su definición la función ln por su deformación lq, presentada en (1), y modificando el cálculo de los valores de expectación a

$$H_q\left[\mathbf{p}_j\right] = -\sum_{i=1}^{k} p_j\left(e_i\right)^q \operatorname{lq}\left(p_j\left(e_i\right)\right) \tag{18}$$

Desde la aparición de esta propuesta de generalización de entropía se han considerado las extensiones de la misma a los distintos campos de aplicación. En particular han surgido generalizaciones de la JSD, que pueden encontrarse resumidas en Martins *et al.*[2] y que pueden asociarse con las interpretaciones antes presentadas.

En una extensión directa como diferencia de entropías, que puede enmarcarse en la interpretación en 2.1, encontramos la propuesta original de Burbea y Rao[6] que define

$$D_q^{(B)}\left[\mathbf{p}_1, \mathbf{p}_2\right] = H_q\left[\mathbf{p}\right] - \sum_{j=1}^{2} \pi\left(j\right) H_q\left[\mathbf{p}_j\right] \tag{19}$$

donde $\mathbf{p}$ y $\mathbf{p}_j$ son respectivamente la distribución de probabilidad marginal definida en (8) y la distribución condicional definida en (10).

A su vez la interpretación como información mutua descripta en 2.2 da lugar a dos generalizaciones. Si consideramos I como la distancia de Kullback-Leibler, siguiendo a Tsallis[7] en su generalización de esta medida de distancia encontramos

$$I_q^{(L)}\left(i,j\right) = \sum_{j=1}^{2}\sum_{i=1}^{k} \frac{\hat{p}\left(e_i,j\right)}{q-1}\left[\left(\frac{\hat{p}\left(e_i,j\right)}{p\left(e_i\right)\pi\left(j\right)}\right)^{q-1} - 1\right] \tag{20}$$

Operando algebraicamente, e identificando I_q con la generalización de la JSD, esta expresión puede escribirse en forma más conveniente como

$$D_q^{(L)}\left[\mathbf{p}_1, \mathbf{p}_2\right] = -\sum_{j=1}^{2}\sum_{i=1}^{k} \pi\left(j\right) p_j\left(e_i\right)^q \times \left[\operatorname{lq}\left(p\left(e_i\right)\right) - \operatorname{lq}\left(p_j\left(e_i\right)\right)\right] \tag{21}$$

según fuera propuesta originalmente por Lamberti y Majtey[8].

Alternativamente, partiendo de la relación

$$I\left(i,j\right) = H\left[\mathbf{p}\right] - H\left[\mathbf{p} \mid \pi\right] \tag{22}$$

donde

$$H\left[\mathbf{p} \mid \pi\right] = -\sum_{j=1}^{2}\sum_{i=1}^{k} \hat{p}\left(e_i,j\right) \ln p_j\left(e_i\right) \tag{23}$$

es la entropía condicional[5] para el símbolo e_i condicionada a la subsecuencia s, Furuichi[9] propone generalizar la información mutua como

$$I_q^{(F)}\left(i,j\right) = H_q\left[\mathbf{p}\right] - H_q\left[\mathbf{p} \mid \pi\right] \tag{24}$$

$$\underbrace{AT \quad \ldots \quad CGT}_{n_1 \text{ símbolos}}\underbrace{TCG \quad \ldots \quad GAT}_{n_2 \text{ símbolos}}$$

Figura 1: Esquema del problema de detección de segmentación: la secuencia $\mathcal{S}$ está conformada por n_1 símbolos con distribución de probabilidades $\mathbf{p}_1$ seguidos de n_2 símbolos con distribución de probabilidades $\mathbf{p}_2$. El problema a resolver es encontrar el valor de n_1 sin el conocimiento de las distribuciones de probabilidad.

generalizando la definición de entropía condicional en forma directa

$$H_q\left[\mathbf{p} \mid \pi\right] = -\sum_{j=1}^{2}\sum_{i=1}^{k} \hat{p}^q\left(e_i,j\right) \operatorname{lq} p_j\left(e_i\right) \tag{25}$$

con lo cual la generalización de la JSD puede expresarse como

$$D_q^{(F)}\left[\mathbf{p}_1, \mathbf{p}_2\right] = -\sum_{i=1}^{k}\left[p\left(e_i\right)^q \operatorname{lq}\left(p\left(e_i\right)\right) - \right. \\ \left. - \sum_{j=1}^{2} \hat{p}^q\left(e_i,j\right) \operatorname{lq}\left(p_j\left(e_i\right)\right)\right] \tag{26}$$

4 Método de segmentación

Consideremos nuevamente una secuencia $\mathcal{S}$ constituída por símbolos e extraídos de un alfabeto $\mathcal{A}$ de k elementos. Supongamos que la secuencia es de composición heterogénea, conformada por dos subsecuencias $\mathcal{S}_1, \mathcal{S}_2$ homogéneas. Denotamos por $f_{i,j}$ la frecuencia de aparición del símbolo e_i en la subsecuencia $\mathcal{S}_j$. Suponemos además que la subsecuencia $\mathcal{S}_j$ tiene largo n_j, con $N = n_1 + n_2$ el largo de la secuencia $\mathcal{S}$, pero suponemos desconocida la posición en la secuencia donde cambia la conformación. El problema a considerar por consiguiente es determinar el punto de segmentación en la secuencia; *i.e.* el valor de n_1. En la figura 1 se ilustra esquemáticamente la conformacin de la secuencia y el valor a determinar.

Supongamos que la conformación de la secuencia está regulada por distribuciones de probabilidad para la aparición de los símbolos, diferentes para cada subsecuencia. El problema a resolver es determinar el punto de segmentación (la posición a partir de la cual cambia la composición) en la secuencia.

Las distribuciones de probabilidad que regulan la conformación de la secuencia no son directamente accesibles. Sólo podemos determinar en forma directa las frecuencias $f_{i,j}$ que dependerán de la posición que se asuma para el punto de segmentación. Podemos así calcular una aproximación a la JSD sustituyendo las distribuciones de probabilidad condicionales, $p_j\left(e_i\right)$ por las frecuencias $f_{i,j}$. En consecuencia los valores que tome D, para un dado punto de segmentación

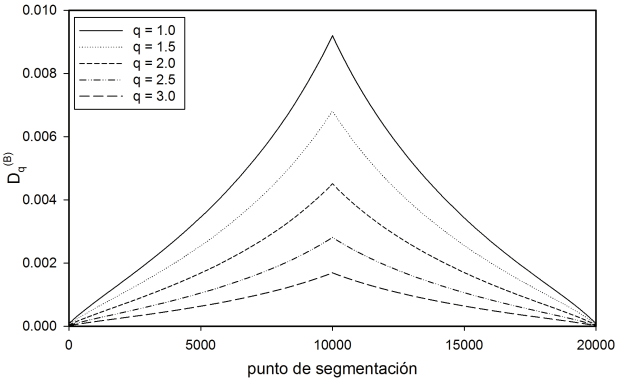

Figura 2: Promedio de $D_q^{(B)}$ sobre 10^4 secuencias formadas por dos subsecuencias de 10^4 elementos cada una. Los valores se han calculado con la generalización de Burbea y Rao (ver texto) para distintos valores del parámetro q. Nótese la disminución en el valor máximo de $D_q^{(B)}$ con el incremento de q.

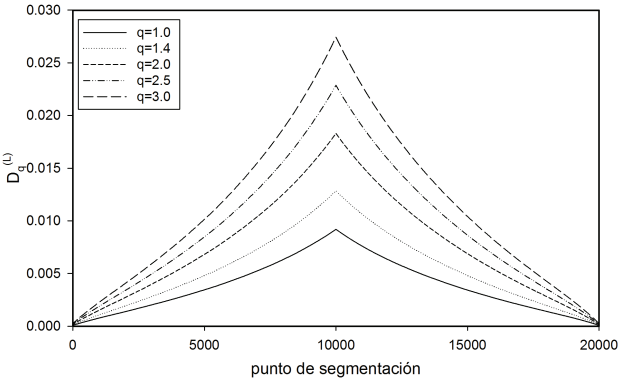

Figura 3: Promedio de $D_q^{(L)}$ sobre 10^4 secuencias formadas por dos subsecuencias de 10^4 elementos cada una. Los valores se han calculado con la generalización de Lamberti y Majtey (ver texto) para distintos valores del parámetro q. Nótese el incremento en el valor máximo de $D_q^{(L)}$ con el incremento de q.

asumido, variarán con cada muestra, aún cuando cada muestra esté generada por la misma distribución de probabilidades.
Un estimador de D, que cuantifique la diferencia entre las dos distribuciones de probabilidad, deberá alcanzar un valor máximo cuando el punto de segmentación elegido para el cálculo de las frecuencias $f_{i,j}$ coincida con la posición en la secuencia donde cambia la composición.
En consecuencia el método de segmentación a proponer consiste en mover un cursor a lo largo de la secuencia $\mathcal{S}$ y calcular el valor de D a partir de las frecuencias $f_{i,j}$ determinadas suponiendo el punto de segmentación en la posición del cursor. El punto de segmentación se toma como la posición del cursor en que D alcanza el valor máximo. Para corroborar esta afirmación y evaluar las posibles mejoras en el método de detección usando generalizaciones no extensivas de la JSD realizamos los experimentos de control siguientes:
Se toma un conjunto de 10^4 secuencias conformadas por dos subsecuencias quiméricas de 10^4 elementos cada una. Se desplaza un cursor a lo largo de cada secuencia así generada, se calcula D para cada posición del cursor y se promedian los valores obtenidos en las distintas secuencias analizadas. Los resultados obtenidos se muestran en las figuras 2 a 5. Se ha calculado D con cada generalización considerada en la sección anterior. En todos los casos el valor $q = 1$ es el determinado por la JSD original definida en (3) a la que se reducen todas las expresiones generalizadas en el límite $q \to 1$.

En la figura 2 presentamos los resultados obtenidos con la generalización en (7) según la propuesta de Burbea y Rao[6]. Como puede verse en esta figura, la posición del punto de segmentación queda bien determinada, aunque el valor del máximo disminuye cuando aumenta q. Dado que el valor de significación[3,?] está determinado por el valor de la JSD en el máximo concluimos que esta generalización a la estadística no extensiva no mejora la probabilidad de detección del punto de segmentación.
En la figura 3 se muestran los resultados obtenidos con la generalización en (23) siguiendo la generalización de Lamberti y Majtey[8]. Puede verse que también en este caso la posición del punto de segmentación queda bien determinada y además el valor del máximo aumenta con q, mostrando por lo tanto una mejora con esta generalización a la estadística no extensiva en la interpretación como información mutua siguiendo la propuesta de Tsallis.

La figura 4 incluye los resultados obtenidos con la generalización en (26) siguiendo la generalización de Furuichi, completando los cálculos con las alternativas

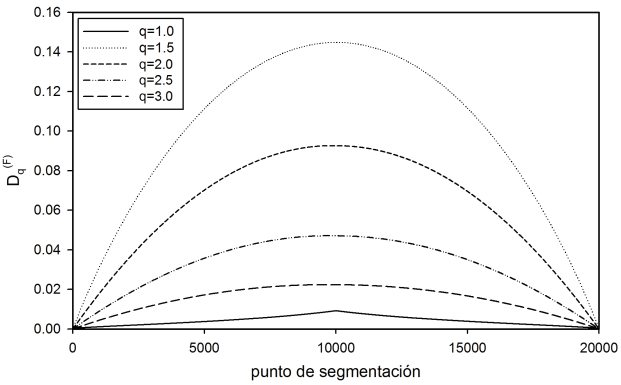

Figura 4: Promedio de $D_q^{(F)}$ sobre 10^4 secuencias formadas por dos subsecuencias de 10^4 elementos cada una. Los valores se han calculado con la generalización de Furuichi (ver texto) para distintos valores del parámetro q. Nótese que aún cuando el valor máximo de $D_q^{(F)}$ crece con el incremento de q, la posición del mismo no queda tan claramente definida como en los casos anteriores.

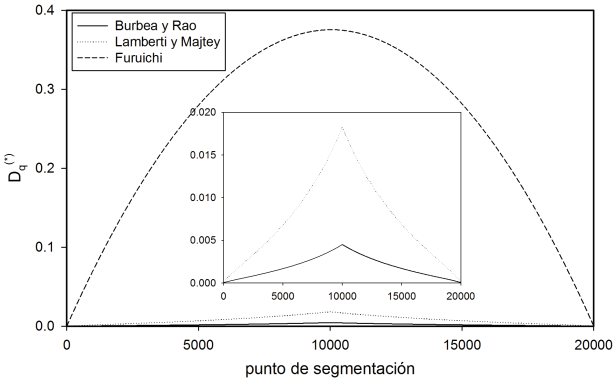

Figura 5: Comparación de los valores promedio de $D_q^{(*)}$ sobre 10^4 secuencias formadas por dos subsecuencias de 10^4 elementos cada una calculados con las distintas generalizaciones consideradas para $q = 2$ (ver texto y figuras anteriores). En el *inset* se han incluido los resultados correspondientes a Burbea y Rao y Lamberti y Majtey en una escala ampliada para una mejor apreciación. Aún cuando los valores de la generalización de Furuichi son claramente mayores que los obtenidos con las restantes generalizaciones, la posición correspondiente al valor máximo no queda tan nítidamente definida.

consideradas. Aún cuando los valores de máximo son mayores para esta generalización, mejorando en principio el criterio de significación, en este caso la posición del punto de segmentación no queda tan claramente definida. Esta indefinición en la posición del máximo hace que la mejora en cuanto al valor del máximo obtenida con esta generalización (información mutua de Furuichi) no facilite la detección del punto de segmentación.

Por último en la figura 5 comparamos los resultados obtenidos con las tres generalizaciones consideradas para un mismo valor $q = 2$. Resulta evidente que la generalización de Furuichi da valores promedio mayores para D_q, en particular para el valor máximo. Sin embargo, como ya fuera señalado, la forma de la curva hace que la posición del valor máximo no quede tan claramente definida como en las otras generalizaciones consideradas. Podemos concluir por lo tanto que, de las alternativas consideradas en la sección 3, la generalización más conveniente para el método de segmentación es la propuesta en[8].

5 Resumen y Conclusiones

Se ha considerado la generalización de la divergencia de Jensen-Shannon (JSD) en el marco de la estadística no extensiva propuesta por Tsallis. En particular se han comparado tres propuestas alternativas que pueden asociarse a las interpretaciones intuitivas presentadas en[2].

La generalización de la JSD, $D_q^{(B)}$, que hemos denominado de Burbea y Rao[6], correspondiente a la sustitución directa de la definición de entropía tradicional de Boltzmann-Gibbs-Shannon por la entropía q de Tsallis en la definición de JSD, presenta una disminución del valor promedio máximo con el incremento de q. Desde este punto de vista esta generalización no resulta conveniente para el método de segmentación.

La segunda generalización considerada, $D_q^{(L)}$, propuesta por Lamberti y Majtey[8] presenta el comportamiento deseado: un aumento en el valor promedio máximo con el incremento de q. El valor máximo determina la posición del punto de segmentación en la secuencia.

La tercera generalización considerada, $D_q^{(F)}$, siguiendo la propuesta de Furuichi[9] presenta también un incremento en el valor promedio máximo con el aumento en el valor de q. Sin embargo, con esta generalización, la posición del punto de segmentación no queda tan netamente definido como en los casos anteriores, si consideramos la forma que adopta la curva.

En conclusión encontramos que de las generalizacions propuestas para la JSD a estadística no extensiva presentes en la literatura, la más apropiada para el método de segmentación resulta la propuesta por Lamberti y Majtey[8].

Esta conclusión reviste importancia no sólo por la utilidad en la aplicación del método de segmentación considerado. Ha sido propuesta recientemene[4] una generalización del método de segmentación basado en modelos markovianos. Resulta de interés analizar la extensión del método en modelos markovianos a la estadística no extensiva de Tsallis. Trabajo en esta línea está en desarrollo y sus resultados se presentarán en futuras comunicaciones.

Agradecimientos: Los autores agradecen el financiamiento de SeCyT-UTN para este proyecto.

Referencias

[1] C. Tsallis, J. Stat. Phys. **52**, 479 (1988).

[2] A. Martins, P. Aguiar y M. Figueiredo, arXiv 0864-1653 (2008).

[3] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán y J. Stanley, Phys. Rev. E bf 65, 041905 (2002).

[4] V. Thakur, R. Azad y R.Ramaswamy, Phys. Rev. E bf 75, 011915 (2007).
A. Arvey, R. Azad y J. Lawrence, Nucelic Acids research, 1 (2009).

[5] T. Cover y J. Thomas, Elements of Information Theory, J. Wiley, New York (1991).

[6] J. Burbea y C. Rao, IEEE Trans. Inf. Theory **28**, 489 (1982).

[7] C. Tsallis, Phys. Rev. E **58**, 1442 (1998).

[8] P. Lamberti y A. Majtey, Phys. A **320**, 81 (2003).

[9] S. Furuichi, J. Math. Phys. **47**, 023302 (2006).