ARTÍCULO DE INVESTIGACIÓN - EDICIÓN ESPECIAL

Vol. 38 | No. 3 | **SEPTIEMBRE - DICIEMBRE 2017** | pp 637-645

REVISTA MEXICANA DE Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.38.3.10

Análisis estadístico de los espectros de frecuencia de las regiones reguladoras del ENCODE

Statistical analysis of the Fourier Spectra of the ENCODE regulatory regions

O. Paredes[1], R. Romo-Vázquez[2], H. Vélez-Pérez[2], J. A. Morales[2]

[1]Maestría en Ciencias en Ingeniería Electrónica y Computación, CUCEI, Universidad de Guadalajara

[2]Departamento de Ciencias Computacionales, CUCEI, Universidad de Guadalajara

RESUMEN

En la actualidad, nuevas bases de datos genómicos (secuencias de ADN) son puestas al alcance del dominio público para su análisis. La bioinformática ha desarrollado algoritmos para extraer información y características de dichas secuencias. Sin embargo, estos algoritmos bioinformáticos tienen limitaciones. Una alternativa es utilizar herramientas propias del procesamiento digital de señales (DSP) adaptadas a secuencias genómicas (procesamiento de señales genómicas - GSP). El presente trabajo versa sobre el análisis de los cuatro primeros momentos centrales (media, desviación estándar, asimetría y curtosis) y dos momentos estadísticos (mediana y varianza) de los espectros frecuenciales de las 15 Regiones Reguladoras (RRs) de la base de datos ENCODE con el objetivo de estudiar diferencias estadísticas y frecuencias características. La base de datos seleccionada es "mapeada". Luego, la FFT es calculada a estas señales genómicas y finalmente los momentos estadísticos son implementados. Los resultados muestran la existencia de 3 grupos de RRs utilizando la media, mediana y curtosis. La desviación estándar y la varianza, parecen no resaltar información importante. Finalmente, la asimetría revela un comportamiento homogéneo ante la presencia de valores atípicos en algunas RRs. Estas observaciones permiten inferir que la periodicidad dentro de la secuencia está relacionada o podría determinar la función biológica que desempeña la misma secuencia.

PALABRAS CLAVE: Procesamiento de señales genómicas, ENCODE, Espectro de frecuencia, Momentos Estadísticos, Transformada de Fourier.

ABSTRACT

Nowadays, new genomic databases (DNA sequences) are available to the whole scientist community for its analysis. The bioinformatics has developed algorithms to extract information and features of the sequences. However, the bioinformatics algorithms have restrictions. An alternative is the use of digital signal processing (DSP) tools adapted to genomic sequences (genomic signal processing - GSP). This work analyzes the first four statistics moments (mean, standard deviation, skewness and kurtosis) and other two moments (median and variance) of the frequency spectra of 15 regulatory regions (RRs) in ENCODE database with the main objective of studying the statistics differences and frequency features. The selected database is mapped. Then, the FFT is calculated to these genomic signals and finally the statistic moments implemented. The results show a three-group behavior in the RRs with the mean, median and kurtosis. The deviations standard and the variance do not show important behavior. Finally, the skewness shows a homogeneous behavior with the lack of atypical values in some RRs. These observations support the idea of the presence of periodicities in a sequence that may be related or may determine the biological function that a sequence may perform.

KEYWORDS: Genomic Signal Processing, ENCODE, Frequency Spectrum, Statistical Moments, Fourier Transform.

Correspondencia

DESTINATARIO: **Omar Paredes**

INSTITUCIÓN: **Centro Universitario de Ciencias Exactas e Ingenierías, Universidad de Guadalajara**

DIRECCIÓN: **Blvd. Marcelino García Barragán #1421, C.P. 44430, Guadalajara, Jalisco, México**

CORREO ELECTRÓNICO: **omar.paredes@alumnos.udg.mx**

Fecha de recepción:

19 de junio de 2017

Fecha de aceptación:

31 de julio de 2017

INTRODUCCIÓN

En la actualidad, nuevas bases de datos genómicos son puestas al dominio público por distintos centros de investigación y laboratorios alrededor del mundo [1]. Lo anterior genera en la comunidad científica pertenecientes al campo de la biología y biomedicina la necesidad de extraer la mayor cantidad de información posible de estas bases de forma rápida y confiable [2-5].

El análisis de la información en una base de datos de secuencias de ADN permite la extracción de características a través de las correlaciones existentes entre las secuencias de ADN, sus periodicidades o en sus *motif* (patrones recurrentes en el ADN [6]), entre otros [7].

En el campo de la bioinformática se han desarrollado diversos algoritmos que resuelven algunas de estas tareas [3, 8]. Sin embargo, la bioinformática presenta ciertas limitaciones para realizar ciertos análisis como es la búsqueda de periodicidad. En este contexto una solución viable es la implementación de herramientas matemáticas propias del procesamiento de señales genómicas [9].

El procesamiento de señales genómicas es un área multidisciplinaria que utiliza herramientas propias del procesamiento digital de señales para extraer información con un significado biológico específico [10,11]

Dentro de este campo, uno de los hallazgos más estudiados e importantes en las regiones codificantes de proteínas es la existencia de un pico en la frecuencia f/3 y la ausencia de este pico en las no codificantes [12, 18].

Sin embargo, las regiones codificantes solo representan el 1% del genoma humano y por varios años el resto del genoma fue considerado ADN "basura" [19, 20]. El proyecto de la Enciclopedia de Elementos del ADN (ENCODE, por sus siglas en inglés) le asignó funciones reguladoras a alrededor del 80% del genoma humano [21-23].

Estas funciones controlan la lectura de la transcripción mediante la promoción, potenciación o silenciando genes, entre otras funciones de regulación [24].

En este trabajo, se presenta un análisis de los momentos estadísticos de los espectros en frecuencia de las regiones reguladoras propuestas por Ernst *et al* [23] que forman parte de la base de datos del ENCODE. La hipótesis de este trabajo es que existen diferencias estadísticas entre los espectros de frecuencia de las 15 regiones reguladoras y que probablemente existan picos de frecuencia característicos como es el caso de las regiones codificantes.

METODOLOGÍA

La metodología que se siguió para el desarrollo de ese trabajo se encuentra esquematizada en la Figura 1.

El primer paso que se realizó fue la elección de los datos. La base de datos que se seleccionó corresponde a los estados de cromatina propuestos por Ernst *et al* [23], la cual se puede consultar en *https://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeBroadHmm* (Junio 2011).

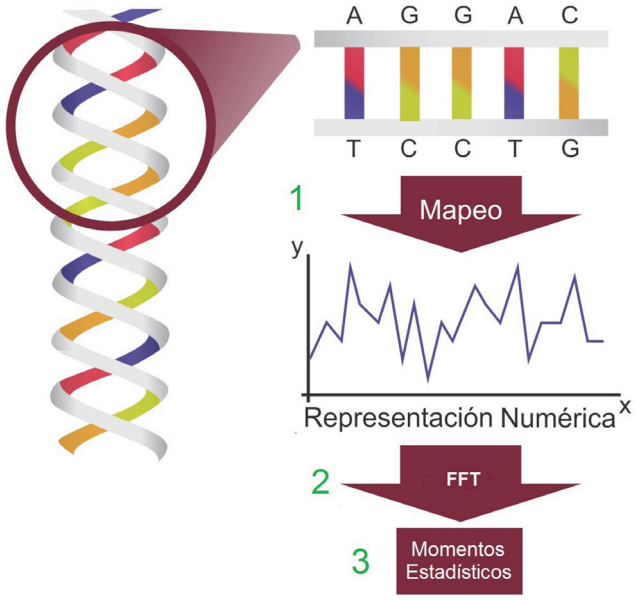

FIGURA 1. Metodología.

Se descargaron los archivos correspondientes a las 15 regiones reguladoras (RRs) (ver Tabla 1) en los 22 autosomas y el cromosoma X en las nueve líneas celulares (CLs) disponibles en la base de datos (ver Tabla 2).

TABLA 1. Lista de las Regiones Reguladoras (RRs). (Tabla reproducida de [23])

Num. de RR	Nombre de Región Reguladora (RR)
1	Active Promoter
2	Weak Promoter
3	Inactived/Poised Promoter
4	Strong Enhancer I
5	Strong Enhancer II
6	Weak/Poised Enhancer I
7	Weak/Poised Enhancer II
8	Insulator
9	Transcriptional Transition
10	Transcriptional Elongation
11	Weak Transcribed
12	Polycomb-repressed
13	Heterochromatin
14	Repetitive I
15	Repetitive I

De acuerdo a la Figura 1, luego de seleccionar una base de datos, el siguiente paso consistió en "mapear" las secuencias. Las secuencias de las RRs fueron extraídas de la construcción del genoma de *http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/referenceSequences/* y posteriormente fueron mapeadas con la técnica de *Neighbor Joining* propuesta por Borrayo *et al* [25].

TABLA 2. Lista de las Líneas Celulares (CLs). (Tabla reproducida de [23])

Num de CL	Abreviatura	Línea celular (CL)
1	GM12878	B-Lymphoblastoid Cells
2	H1esc	Embryonic Stem Cells
3	Hepg2	Hepatocellular Carcinoma Cells
4	Hmec	Mammary Epithelial Cells
5	Hsmm	Skeletal Muscle Myoblasts
6	Huevec	Umbilical Vein Endothelial Cells
7	K562	Erytrocytic Leukimia Cells
8	Nhlf	Normal Lung Fibroblasts
9	Nhek	Normal Epidermal Keratinocytes

Se entiende por mapeo el proceso de trasladar la información contenida en las secuencias de ADN (secuencias ordenadas de los nucleótidos, representados por las letras A, T, C, G) a una representación numérica (señal genómica) con la menor pérdida de información [7]. El mapeo de *Neighbor Joining* [25] consiste en dos etapas, la primera es explicada por la Ecuación 1:

$$(S_i, S_{i+1}) \rightarrow \hat{S}_i \qquad (1)$$

para i = 1,2,3,..., l donde l es la longitud de la secuencia de ADN y $S_i \in \{A,T,C,G\}$ que a su vez representan los nucleótidos adenina (A), timina (T), citosina (C), guanina (G); $\hat{S}_i \in \{x_1,x_2,...,x_{16}\}$ y los valores de $x_{1,2,...,16}$ son enteros equidistantes, el nuevo vector $\hat{S}$ tiene una longitud de l - 1.

La segunda etapa del mapeo consiste en realizar un ventaneo del vector $\hat{S}$ y es calculada por la Ecuación 2:

$$\tilde{S}_i = \frac{1}{2\alpha+1} \sum_{\tau=i}^{2\alpha+i} \hat{S}_\tau \qquad (2)$$

donde $\hat{S}$ es el vector resultado de la Ecuación 1, $\tilde{S}$ es un nuevo vector de longitud l - 2α y α es un número entero que representa la cantidad de nucleótidos considerados vecinos a ambos lados del nucleótido evaluado. En este trabajo se usó el valor α=3 y los valores de $x_{1,2,...,16}$ propuestos por Borrayo *et al* [25].

Una vez mapeadas las secuencias a señales genómicas, se procedió a calcular la transformada de Fourier de cada una de ellas. La transformada de Fourier (ver Ecuación 3) es una herramienta matemática que permite calcular la periodicidad de una señal [9].

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt \qquad (3)$$

A continuación, agrupamos los espectros de las señales genómicas en 3105 conjuntos donde cada conjunto representa las señales genómicas que cumplen una de las RRs de las enlistadas en la Tabla 1, pertenecen a una LC de las presentadas en la Tabla 2 y a uno de los 23 cromosomas analizados.

La magnitud de cada coeficiente de Fourier (valor absoluto) fue calculada para obtener números reales y reducir la complejidad del análisis. Después se realizó una normalización de los datos con el objetivo de realizar una comparación de los mismos sin que la cantidad de puntos pudieran producir ruidos o valores atípicos. Los datos fueron normalizados entonces en función de 3 criterios: la longitud, la energía y la banda de frecuencia:

i. Normalización por longitud: consiste en interpolar lineal cada espectro de Fourier a un valor de 800 puntos que representa la mediana de tamaño del conjunto total de señales genómicas del estudio.

ii. Normalización por energía: consiste en dividir el espectro normalizado por longitud entre el valor del tamaño original del espectro.

iii. Normalización por banda de frecuencia: recordando que las señales genómicas fueron agrupadas en conjuntos que pertenecen a la misma RR, CL y cromosoma. Para cada conjunto las bandas de frecuencia (800) son divididas por el valor máximo de la banda evaluada.

Luego de realizar las normalizaciones pertinentes, se procedió de acuerdo al paso 3 de la Figura 1 al cálculo de momentos estadísticos. Sobre los espectros de frecuencia se calcularon los cuatro primeros momentos centrales (media, desviación estándar, asimetría y curtosis), y dos momentos estadísticos más (mediana y varianza).

La media es el primer momento estadístico central y es considerada el valor representativo de la distribución de coeficientes el cual puede ser interpretado como el valor medio de energía, donde un valor mayor representa coeficientes con una organización posible-

mente definida y no aleatoria. La mediana se refiere al valor medio de energía de la distribución y puede ser semejante a la media si es una distribución normal.

La desviación estándar y varianza son medidas estadísticas de dispersión y que nos describen la tendencia a mantener el orden (consideramos orden a la cualidad de que existan picos frecuenciales que sobresalen sobre el resto). Una baja desviación estándar, es decir dispersión baja, corresponde a la posible presencia de picos frecuenciales distinguibles.

La asimetría es el tercer momento estadístico central y nos describe las colas en una distribución lo que está relacionado con la existencia de valores atípicos. La curtosis describe la persistencia a concentrarse los datos de la distribución en el valor medio de energía, lo cuál reforzaría la idea de la presencia de picos frecuenciales representativos en los espectros. Las ecuaciones de los momentos estadísticos descritos se encuentran en la Tabla 3.

TABLA 3. Ecuaciones correspondientes a los momentos estadísticos utilizados.

Momento Estadístico	Ecuación
Media ($\bar{x}$)	$\frac{\sum_{i=1}^{n} x_i}{n}$
Desviación Estándar (σ)	$\frac{\sqrt{\sum_{i=1}^{n}(\bar{x}-x_i)^2}}{n}$
Asimetría ($skew$)	$\frac{\frac{\sum_{i=1}^{n}(\bar{x}-x_i)^3}{n}}{\left[\frac{\sum_{i=1}^{n}(\bar{x}-x_i)^2}{n-1}\right]^{3/2}}$
Curtosis ($kurt$)	$\frac{\frac{\sum_{i=1}^{n}(\bar{x}-x_i)^4}{n}}{\left[\frac{\sum_{i=1}^{n}(\bar{x}-x_i)^2}{n-1}\right]^{2}} - 3$
Mediana (M_e)	Si es impar $M_e = x_{n+1/2}$ Si es par $M_e = {(x_{n/2} + x_{n+1/2})}/{2}$
Varianza (σ^2)	$\frac{\sum_{i=1}^{n}(\bar{x}-x_i)^2}{n}$

RESULTADOS Y DISCUSIÓN

El objetivo de este trabajo es encontrar las diferencias estadísticas entre los espectros de frecuencia de las 15 regiones reguladoras. Para conseguir esto se calcularon seis momentos estadísticos: media ($\bar{x}$), desviación estándar (σ), asimetría (*skew*), curtosis (*kurt*), mediana (M_e) y varianza (σ^2).

En la Figura 2 se presentan los momentos estadísticos de la LC GM12878 en todas las RRs y cromosomas. Se puede observar la existencia de 3 grupos de regiones reguladoras en la media, mediana y curtosis. Esto es interpretado como grupos de RRs con secuencias con un grado de orden (tendencia a estar ordenados y no dispuestos aleatorios) determinado.

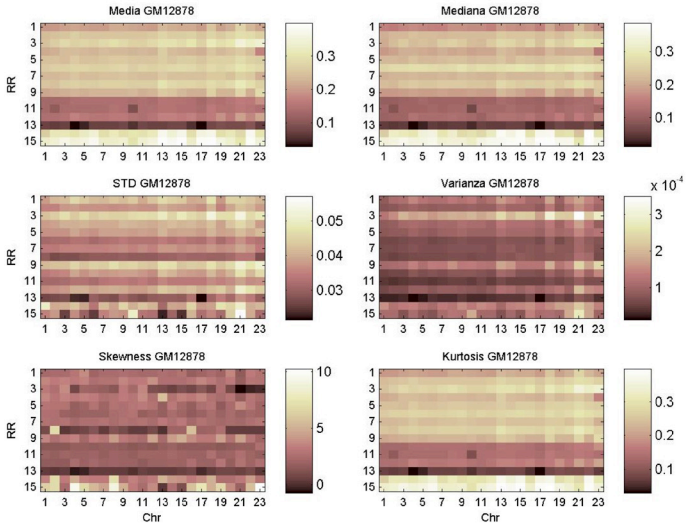

FIGURA 2. Momentos estadísticos de la línea celular GM12878.

El primer grupo está conformado por las RRs Repetitive I y Repetitive II interpretado como RRs altamente ordenadas. Por el contrario el grupo de RRs con bajo orden o alta aleatoriedad está conformado por las RRs *Transcriptional Elongation*, *Weak Transcribed*, *Polycomb-repressed*, *Heterochromatin*. Esta observación es congruente con el principio de mínimo esfuerzo donde secuencias largas, como lo son las pertenecientes a estas RRs, implican un gasto energético muy alto.

El resto de RRs están medianamente ordenadas (refiriéndonos a las secuencias de ADN).

Con respecto a la desviación estándar y la varianza, visualmente parecen no contener información relevante. Sin embargo, observando el rango en las que oscilan (0 - 0.05), podemos inferir que el orden, ya sea alto, medio o bajo, es preservado debido a que se puede considerar ambos valores de los momentos estadísticos como bajos. El asimetría nos muestra un comportamiento homogéneo en la presencia de valores atípicos en las primeras 13 RRs. Las *Repetitive* I y II tienen valores de asimetría mayores, es decir, que no todas las secuencias de ADN pertenecientes a estas RRs son altamente ordenadas. Este comportamiento se puede observar a través de las otras ocho LCs (ver figuras suplementarias S24-S31*).

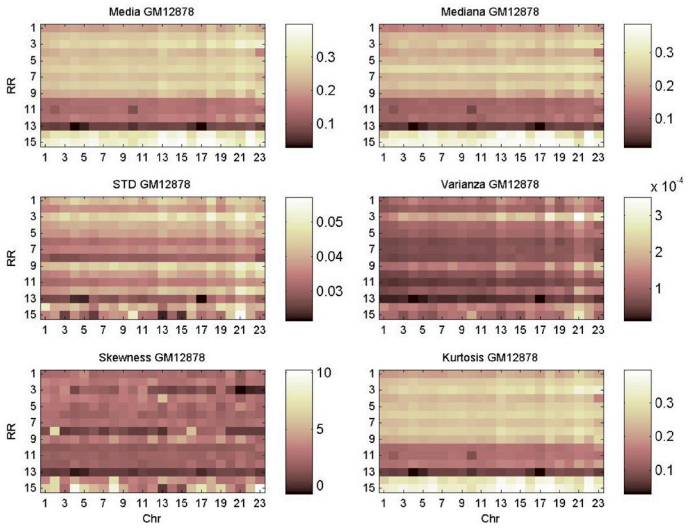

FIGURA 3. Momentos estadísticos del cromosoma 4.

El comportamiento descrito anteriormente puede ser observado en la Figura 3, donde se analizan los datos pertenecientes a un cromosoma a través de todas las RRs y las LCs. El resto de los cromosomas se pueden ver en las figuras suplementarias S1-S23*. En todas ellas se refuerza la observación descrita.

Analizando una única RR en todas las LCs y cromosomas, por ejemplo *Active Promoter* (ver Figura 4 y Figura 5) se puede observar que un comportamiento en la LC 2 (H1esc) es distinto al resto de las LCs. Las secuencias que son *Active Promoter* son las encargadas de promover la expresión de genes. Por el contrario, las secuencias con la función de *Poised Promoter* se encargan de reprimir o silenciar la expresión de los genes. Estas características son descritas en [26-28] y corresponden a las de las células madres.

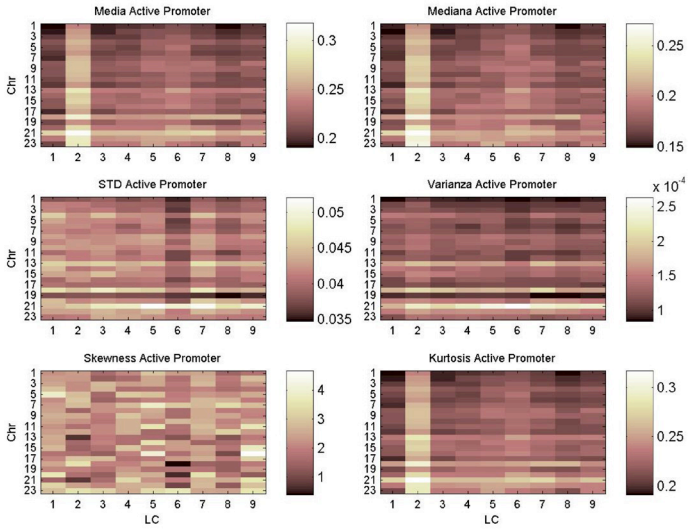

FIGURA 4. Momentos estadísticos de la región reguladora Active Promoter.

Los valores altos en el caso de la RR *Active Promoter* en la media, en la mediana y en las curtosis son significativos, porque los promotores representan secuencias altamente ordenadas y, probablemente, accesadas continuamente.

En el caso contrario, de los bajos valores de los mismos momentos en la RR *Poised Promoter* se puede inferir que son altamente desordenadas. Esto es relevante porque son secuencias no necesarias en las células madres. Revisitando la idea de que una célula madre expresa la mayoría de los genes y silencia pocos genes, caso contrario de células diferenciadas como las otras 8 LCs, que expresa una cantidad pequeña de genes y silencia una gran cantidad [26-28].

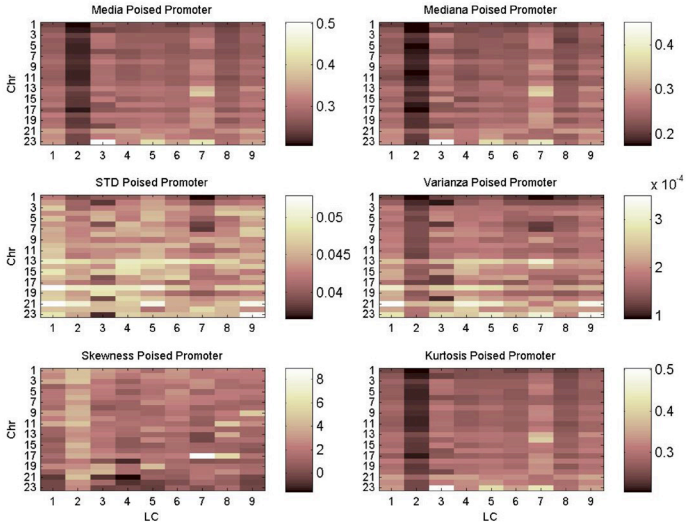

FIGURA 5. Momentos estadísticos de la región reguladora Poised Promoter.

El comportamiento de los momentos estadísticos de nuestro trabajo y la descripción dada por la literatura es relevante porque conecta el comportamiento frecuencial de las secuencias con características biológicas no determinadas con esta metodología lo cual da un argumento a favor para el uso del procesamiento de señales como una herramienta para análisis de datos genómicos en la disciplina del procesamiento de señales genómicas.

Como una línea de investigación futura a este trabajo puede implementarse una etapa de clasificación o clustering extrayendo una cantidad mayor de información que está contenida dentro de los espectros de frecuencia y no puede ser abordada solamente con índices estadísticos.

CONCLUSIONES

El análisis de datos genómicos es crucial dado que la comunidad científica deja al alcance del dominio público cada vez más bases de datos. Sin embargo poca información es extraída de estas. El procesamiento de señales genómicas es una disciplina poco explorada hasta el momento, aunque ha dado resultados interesantes como el descubrimiento de la frecuencia f/3 en la secuencias codificantes.

En este trabajo, hemos implementado el procesamiento de señales genómicas para analizar estadísticamente las frecuencias de secuencias no codificantes con 15 funciones biológicas específicas.

Los resultados obtenidos en este trabajo reflejan que los espectros de frecuencia de las señales son distintos entre funciones biológicas y que ciertas características biológicas, como en el caso de la línea celular H1esc, se pueden analizar con esta metodología.

Los resultados de este trabajo nos han permitido comprobar que el procesamiento de señales genómicas es una herramienta poderosa para la extracción de características. Además, y da pauta a que más herramientas nativas del procesamiento de señales puedan ser adoptadas a las señales genómicas para obtener información biológica relevante y que aún permanece oculta con los métodos tradicionales de bioinformática y genética.

***Este artículo contiene un anexo de imágenes, el cual puede ser consultado en: https://goo.gl/5LEf6r**

REFERENCIAS

[1] Grivell L. Mining the bibliome: searching for a needle in a haystack?: New computing tools are needed to effectively scan the growing amount of scientific literature for useful information. EMBO Rep [Internet]. 2002 Mar 1;3(3):200-3. Available from: http://embor.embopress.org/cgi/doi/10.1093/embo-reports/kvf059

[2] Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, et al. The Ontology for Biomedical Investigations. Xue Y, editor. PLoS One [Internet]. 2016 Apr 29;11(4):e0154556. Available from: http://dx.plos.org/10.1371/journal.pone.0154556

[3] Fuchs R. From Sequence to Biology: The Impact on Bioinformatics. Bioinformatics [Internet]. 2002 Apr 1;18(4):505-6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12016047

[4] Yu U, Lee S-H, Kim Y-J, Kim S. Bioinformatics in the Post-genome Era. BMB Rep [Internet]. 2004 Jan 31;37(1):75-82. Available from: http://koreascience.or.kr/journal/view.jsp?kj=E1MB-B7&py=2004&vnc=v37n1&sp=75

[5] Kanehisa M, Bork P. Bioinformatics in the post-sequence era. Nat Genet [Internet]. 2003 Mar;33(3s):305-10. Available from: http://www.nature.com/doifinder/10.1038/ng1109

[6] D'haeseleer P. What are DNA sequence motifs? Nat Biotechnol [Internet]. 2006 Apr;24(4):423-5. Available from: http://dx.doi.org/10.1038/nbt0406-423

[7] Ahmad M, Jung LT, Bhuiyan A-A. From DNA to protein: Why genetic code context of nucleotides for DNA signal processing? A review. Biomed Signal Process Control [Internet]. 2017 Apr;34:44-63. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1746809417300046

[8] Gibson TA. The Roots of Bioinformatics. PLoS Comput Biol [Internet]. 2012 Aug 30;8(8):e1002679. Available from: http://dx.plos.org/10.1371/journal.pcbi.1002679

[9] Afreixo V, Ferreira PJSG, Santos D. Fourier analysis of symbolic data: A brief review. Digit Signal Process [Internet]. 2004 Nov;14(6):523-30. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1051200404000466

[10] Dougherty E, Cai X, Huang Y, Kim S, Yamaguchi R. Editorial [Hot topic: Genomic Signal Processing: Part 1 (Guest Editors: E.R. Dougherty, X. Cai, Y. Huang, S. Kim and R. Yamaguchi)]. Curr Genomics [Internet]. 2009 Sep 1;10(6):364-364. Available from: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1389-2029&volume=10&issue=6&spage=364

[11] Anastassiou D. Genomic signal processing. IEEE Signal Process Mag [Internet]. 2001 Jul;18(4):8-20. Available from: http://ieeexplore.ieee.org/document/939833/

[12] Fickett JW. Recognition of protein coding regions in DNA sequences. Nucleic Acids Res [Internet]. 1982;10(17):5303-18. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/10.17.5303

[13] Yin C, Yau SS-T. A Fourier Characteristic of Coding Sequences: Origins and a Non-Fourier Approximation. J Comput Biol [Internet]. 2005 Nov;12(9):1153-65. Available from: http://online.liebertpub.com/doi/pdf/10.1089/cmb.2005.12.1153%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/16305326

[14] Wang L, Stein LD. Localizing triplet periodicity in DNA and cDNA sequences. BMC Bioinformatics [Internet]. 2010;11(1):550. Available from: http://www.biomedcentral.com/1471-2105/11/550

[15] Yin C, Yau SS-T. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. J Theor Biol [Internet]. 2007 Aug;247(4):687-94. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0022519307001543

[16] Eskesen ST, Eskesen FN, Kinghorn B, Ruvinsky A. Periodicity of 6DNA in exons. BMC Mol Biol [Internet]. 2004;5(1):12. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=516030&tool=pmcentrez&rendertype=abstract

[17] Trotta E. The 3-Base Periodicity and Codon Usage of Coding Sequences Are Correlated with Gene Expression at the Level of Transcription Elongation. Kudla G, editor. PLoS One [Internet]. 2011 Jun 28;6(6):e21590. Available from: http://dx.plos.org/10.1371/journal.pone.0021590

[18] Chechetkin VR, Turygin AY. Size-dependence of three-periodicity and long-range correlations in DNA sequences. Phys Lett A [Internet]. 1995 Mar;199(1-2):75-80. Available from: http://linkinghub.elsevier.com/retrieve/pii/0375960195000477

[19] Human Genome Sequencing Consortium I. Finishing the euchromatic sequence of the human genome. Nature [Internet]. 2004 Oct 21;431(7011):931-45. Available from: http://www.nature.com/doifinder/10.1038/nature03001

[20] Palazzo AF, Gregory TR. The Case for Junk DNA. Akey JM, editor. PLoS Genet [Internet]. 2014 May 8;10(5):e1004351. Available from: http://dx.plos.org/10.1371/journal.pgen.1004351

[21] Qu H, Fang X. A Brief Review on the Human Encyclopedia of DNA Elements (ENCODE) Project. Genomics Proteomics Bioinformatics [Internet]. 2013 Jun;11(3):135-41. Available from: http://dx.doi.org/10.1016/j.gpb.2013.05.001

[22] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature [Internet]. 2012 Sep 5;489(7414):57-74. Available from: https://goo.gl/nmb68z

[23] Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature [Internet]. 2011 May 5;473(7345):43-9. Available from: http://dx.doi.org/10.1038/nature09906

[24] Maston G a, Evans SK, Green MR. Transcriptional Regulatory Elements in the Human Genome. Annu Rev Genomics Hum Genet [Internet]. 2006 Sep;7(1):29-59. Available from: http://www.annualreviews.org/doi/10.1146/annurev.genom.7.080505.115623

[25] Borrayo E, Mendizabal-Ruiz EG, Vélez-Pérez H, Romo-Vázquez R, Mendizabal AP, Morales JA. Genomic Signal Processing Methods for Computation of Alignment-Free Distances from DNA Sequences. Bajic VB, editor. PLoS One [Internet]. 2014 Nov 13;9(11):e110954. Available from: http://dx.plos.org/10.1371/journal.pone.0110954

[26] Ong CT, Corces VG. Enhancers: emerging roles in cell fate specification. EMBO Rep [Internet]. 2012;13(5):423-30. Available from: https://goo.gl/ufyvEv

[27] Kang L, Gao S. Pluripotency of induced pluripotent stem cells. J Anim Sci Biotechnol [Internet]. 2012;3(1):5. Available from: https://goo.gl/9RaFgH

[28] Jaenisch R, Young R. Stem Cells, the Molecular Circuitry of Pluripotency and Nuclear Reprogramming. Cell. 2008;132(4):567-82.