Firmas Genéticas en secuencias de ADN: Un análisis en Regiones Codificantes y no Codificantes de Proteínas

Rosario Medina Rodriguez, Jesus Mena-Chalco

Resumen — La representación de genomas completos, compuestos por millones de nucleótidos, usando estructuras genómicas o componentes menores ha sido objeto de atención en los últimos años pues se acredita fuertemente que toda especie biológica existente puede ser representada por una "Firma Genética". Podemos interpretar la denominada firma genética como un conjunto de medidas, dependientes de resolución o granularidad, que intrínsecamente representa la organización primaria de una secuencia de ADN de un organismo. Hoy en día, en el área de Bioinformática a nivel de nucleótidos, la prioridad es obtener la mayor cantidad de información posible de cada genoma secuenciado, con la finalidad de conseguir un mejor entendimiento de la taxonomía y composición genómica de los organismos. En el presente artículo, correspondiente al tema de tesis de pre-grado aún en desarrollo, se describe una breve introducción a las firmas genéticas usando *Chaos Game Representation of Frequencies* (FCGR) y como parte inédito se evalúa la influencia que tienen las regiones codificantes y no codificantes de proteínas en la representación de genomas, realizada en este caso, a través de firmas genéticas.

Términos de indexación — Regiones Codificantes, Regiones no Codificantes, Firmas Genéticas, FCGR.

1. Introducción

La reciente disponibilidad de largas secuencias genómicas abren un nuevo campo de investigación dedicado al análisis de su estructura. (Beutleretal.1989;Woese,Kandler,y Wheelis 1990; Charlesworth 1994; Sharpand Matassi 1994;Doolittle 1997; Maleyand Marshall 1998).

Actualmente existe una gran cantidad de organismos con secuencias de ADN de genomas completos conocidos y almacenados en repositorios de datos genéticos para su posterior análisis[1].

Los genomas al estar conformados por millones de nucleótidos, hacen que los análisis sobre los mismos resulten complejos y a veces no realizables debido a limitaciones computacionales (como de memoria y procesamiento). Por ese motivo en la comunidad de Bioinformática existe la preocupación de encontrar una forma de caracterización de genomas que permita realizar una representación de los mismos de manera que muestren sus características principales (con una reducción en su dimensionalidad). Así, una forma de caracterizar un genoma es conocida como "firma genética".

El trabajo de tesis, aún en desarrollo, se esta orientando al estudio inédito de la influencia de las regiones genómicas (regiones codificantes y regiones no codificantes) en las firmas genéticas, es decir, que regiones contienen mayor cantidad de información relevante en la secuencia de ADN, para caracterizar al genoma; permitiendo así reducir aun mas la dimensionalidad y a la vez el costo de almacenamiento y análisis de las mismas, al usar sólo las regiones más importantes.

El resto del artículo esta organizado de la siguiente manera: En la sección 2 son tratados algunos conceptos básicos sobre el trabajo de manera que sea de rápido entendimiento. En la sección 3, se verán algunos trabajos previos que justifiquen el presente. La técnica a usar para representar los genomas, será vista en la sección 4. El conjunto de datos usados en el trabajo esta descrito en la sección 5. Algunas técnicas de agrupamiento en la sección 6. Finalmente, en las secciones 7 y 8 son mostrados los resultados obtenidos, su apropiada disposición y conclusiones, respectivamente.

2. Conceptos Básicos

2.1. Firmas Genéticas

Una firma genética esta asociada con un orden particular o una longitud de subsecuencia que representa una medida de la resolución o granularidad en el análisis de la organización primaria de una secuencia de ADN, según [13].

2.2. Regiones Codificantes y No codificantes

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) [9].

Rosario Medina R. es estudiante de Quinto Año de la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional de San Agustín, Arequipa. Perú. E-mail: rosario1316@gmail.com

Jesus Mena-Chalco es estudiante de Doctorado de la Universidad de São Paulo (IME-USP). Brasil. E-mail: jmena@vision.ime.usp.br

[1]Uno de los repositorios genéticos ampliamente conocido es el perteneciente al *National Center for Biotechnology Information (NCBI)*: www.ncbi.nlm.nih.gov/

La Región para la codificación de Proteínas (CDS), para organismos procariontes se considera una única región, entre tanto para organismos eucariontes es considerada una secuencia alternada de exones/intrones separadas por regiones de corte y regiones aceptadoras.

Se considera a un ***exón*** como una región necesaria para la codificación de proteínas. De se mismo modo, se considera a un ***intrón*** como una región no presente en la codificación, a menudo extensas y con funciones aún desconocidas [2].

3. Trabajos Previos

Entre las diferentes formas de representar secuencias de ADN, tenemos:

- *Chaos Game Representation* (CGR) fue propuesta como una representación independiente de la escala para secuencias genómicas por Jeffrey en 1990 [7]. La técnica, formalmente un mapa interactivo, puede remontarse aun mas atrás, a los fundamentos de la mecánica estadística, en particular a la teoría del Caos (Bar-Yam, 1997).

  El espacio CGR es un sistema contínuo de referencia, donde todas las posibles secuencias de cualquier longitud tienen una única posición. Consecuentemente toda posible sucesión de nucleótidos será codificada en un espacio continuo [1].

- Sin embargo en [4], se desarrolla una versión del método de CGR, [1] propuso el nombre FCGR(matrices de frecuencia extraídas de CGR) para nombrar esta modificación de *Chaos Game Representation*, que permite la cuantificación de patrones observados y un procesamiento rápido de secuencias muy largas.

En [7] se propuso la representación de genomas usando FCGR, se obtuvieron firmas genéticas, para la región del cromosoma 11 de la secuencia *Human Beta Globin* (HUMHBB), de este trabajo se toma en cuenta la idea propuesta como *Preguntas Abiertas*: .[Es] posible aplicar el algoritmo de CGR en codones (regiones codificantes de proteínas)o a los aminoácidos que estos codifican...''

En [4] se realizaron las firmas genéticas de siete especies usando FCGR, luego se aplico un analisis de componentes principales (PCA, *Principal Component Analysis*) y posteriormente se estableció la distancia que había entre cada firma, determinando así similitudes filogenéticas entre las especies analizadas.

4. Chaos Game Representation of Frequencies - FCGR

Las secuencias genómicas están en un constante estado de variación debido a procesos, tales como la transposición, transformación, translocation y recombinación. (Karlin et al, 1998; Casjens,1998)

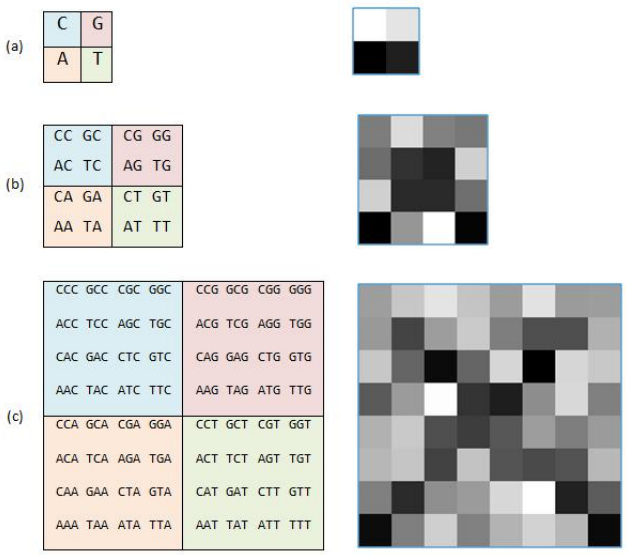


Figura 1: Configuración de frecuencias (columna izquierda) y firma genética (columna derecha) para tamaños de oliginucleótidos (a) de longitud 1, (b) de longitud 2, y (c) de longitud 3. Para las firmas genéticas fue usado el genoma completo de *Archaeglobus Fulgidus*.

Básicamente, todo el conjunto de frecuencias de oligonucleótidos, encontrados en una secuencia genómica dada, pueden ser mostrados en la forma de una sola imagen en la cual cada pixel está asociado a una cadena de oligonucleótidos específica. Las frecuencias encontradas en una secuencia, son mostradas en una imagen cuadrada y la posición de cada secuencia de oligonucleótidos es escogida de acuerdo a un procedimiento recursivo. Es por eso que la imagen es dividida en cuatro cuadrantes en las cuales, las secuencias que terminan en una base apropiada son recolectadas.

En FCGR la imagen es divida en 4^n cuadrados, donde n es la longitud de los oligonucleótidos a representar. Para cada oligonucleótido un FCGR debe ser generado. En la figura 1 obsérvese tres ejemplos de configuración de frecuencias de oligonucleótidos. La firma genética obtenida En la Figura 1(a) corresponde simplemente a la frecuencia absoluta de los nucleótidos A, C, G e T. Es importante destacar que, la frecuencia de oligonucleótidos es representada por una escala de grises, siendo que la mayor probabilidad esta representada por el color negro. Entretanto, la firma genética para tamaño de oligonucleótido 2, mostrada en la figura 1(b), las frecuencias consideradas serán las correspondientes a la probabilidad de encontrar las siguientes palabras en el genoma: AA, CA, GA, TA, AC, CC, GC, TC, AG, CG, GG, TG, AT, CT, GT, y TT.

En este contexto, cuando lo que se busca es representar tetranucleótidos en una imagen FCGR, un total de 4^4 (256) pequeños cuadrados formarán la imagen completa; cada pequeño cuadrado corresponde a un tetranucleótido, como se ve en la figura 2.

En [4] nos dice; usando CGR, se observa que las subse-

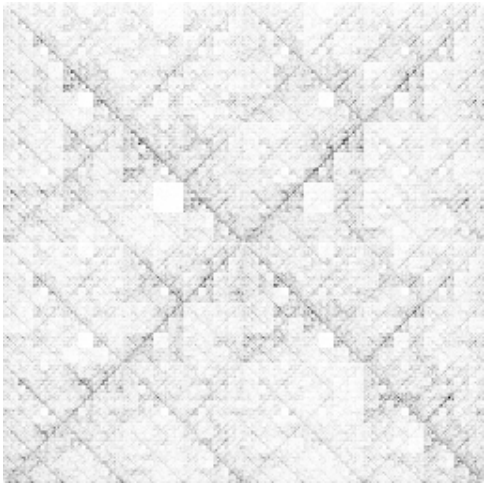


Figura 2: FCGR para *Archaeglobus Fulgidus*.

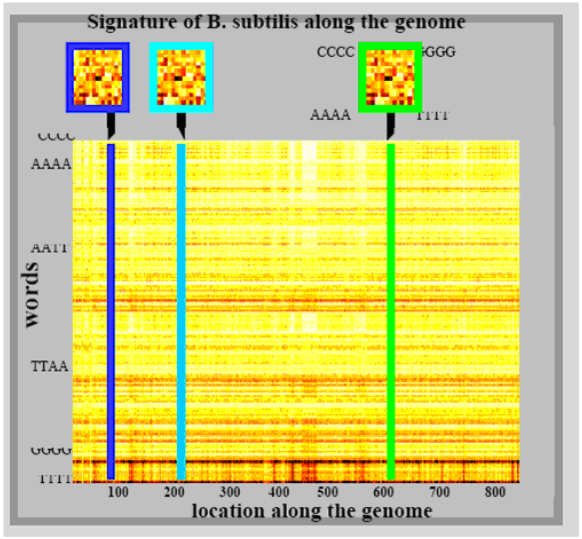


Figura 3: Firma genética aplicando FCGR de *Bacillus subtitlis* cada 3000 nucleótidos.

cuencias de un genoma, muestran las principales características de todo el genoma, comprobando así la validez de la firma genómica.

La firma genética es especie-específica, según [11]. El estudio de oligonucleótidos (o palabras) encontrados en genomas debería ayudar a detectar factores de especificidad. El uso de fragmentos de mediano tamaño (5 a 10 kb) permite una casi perfecta clasificación y puede incluso ser usada para diferenciar especies muy similares.

Así como en [3] se afirma que, la firma genética expresa el uso de pequeñas cadenas de oligonucleótidos en una secuencia. Esta se puede mostrar como una imagen, donde cada cuadrado representa la frecuencia de una palabra dada. En ese estudio, el genoma *Bacillus subtitlis* es escaneado a través de ventanas de 3000 nucleótidos (firmas locales). Firmas de ventanas sucesivas son mostradas como líneas verticales consecutivas. La firma genómica (con ligeras variaciones) es observable en muchas de las ventanas, como en la figura 3. Desde que una figura es observable en todas las especies, la invarianza de la firma a lo largo del genoma lleva a un estilo *especie-específico*.

5. Conjuntos de datos

Para validar nuestro trabajo consideraremos datos correspondientes 21 especies, obtenidas del repositorio NCBI-GenBank, cuya taxonomía esta detallada en la tabla 1[2]:

De todas las especies consideradas, fueron creados 4 conjuntos de prueba para nuestros experimentos de firmas genéticas:

1. Genoma Completo para cada especie.
2. 50 subsecuencias de tamaño 100 000 para cada especie. Siendo que fueron aleatoriamente extraidas 1000 cadenas de tamaño 100.

[2]Taxonomía obtenida de *NCBI Taxonomy Database*: http://www.ncbi.nlm.nih.gov/Taxonomy/

Especie	Reino	Genero	Tamaño
A. fulgidus	Archaea	Archaeoglobus	2.158 Kb
B. burgdorferi	Bacteria	Borrelia	31 Kb
C. acetobutylicum	Bactera	Clostridium	3.904 Kb
V. cholerae	Bacteria	Vibrio	4.711 Kb
E. coli	Bacteria	Escherichia	4.596 Kb
A. fumigatus	Fungi	Aspergillus	4.873 Kb
C. albicans	Fungi	Candida	941 Kb
E. cuniculi	Fungi	Encephalitozoon	209 Kb
E. gossypii	Fungi	Eremothecium	686 Kb
M. jannaschii	Bacteria	Marinobacterium	1.659 Kb
M. leprae	Bacteria	Mycobacterium	3.238 Kb
T. maritima	Bacteria	Thermotoga	1.844 Kb
D. melanogaster	Animalia	Drosophila	1.256 Kb
M. tuberculosis	Bacteria	Mycobacterium	4.363 Kb
T. pallidum	Bacteria	Treponema	1.128 Kb
S. pneumoniae	Bacteria	Streptococcus	2.020 Kb
D. radiodurans	Bacteria	Deinococcus	2.624 Kb
S. solfataricus	Archaea	Sulfolobus	2.964 Kb
S. sp PCC6803	Bacteria	Synechocystis	3.540 Kb
A. tumefaciens	Bacteria	Agrobacterium	2.815 Kb
B. subtilis	Bacteria	Bacillus	4.175 Kb

Cuadro 1: Características de las especies consideradas en nuestro trabajo.

3. Subsecuencia de tamaño 100 000 extraída de las Regiones Codificantes, de cada especie.
4. Subsecuencia de tamaño 100 000 extraída de las Regiones No Codificantes, de cada especie.

Cabe destacar que:

- Las firmas genéticas fueron obtenidas a través de la técnica FCGR para oligonucleótidos de tamaño 8, obteniendo entonces imágenes de $2^8 X 2^8$ pixels, en

un total de 65 536 bases.

- Las Regiones Codificantes y No Codificantes de los genomas, fueron extraídas usando el método basado la MMT (*Transformada Modificada de Morlet*) descrito en [9], compuesto por tres pasos:
  - Mapeamiento numérico de una secuencia de ADN a cuatro secuencias binarias.
  - Aplicación de la MMT a cada secuencia binaria.
  - Proyección de las secuencias espectrales sobre el eje de las posiciones.
  - Extracción de 100 000 bases correspondientes a regiones cuyos coeficientes de proyección fueron mayores al 80 % del valor de proyección más grande obtenido en el análisis.

6. Técnicas de agrupamiento

Según [6], *Clustering* es una clasificación de patrones no supervisada en grupos (*clusters*). Existen diferentes técnicas de agrupamiento de datos, entre ellas tenemos:

1. ***Least Square Projection*(LSP) :** [5]
   Dado un conjunto de puntos $S = \{p_1, ..., p_n\}$ en R^m, el algoritmo LSP tiene como objetivo representar los puntos de S es un espacio de menor dimension R^d, $d < m$, de manera que se preserve la relación de vecindad entre los puntos tanto como sea posible.

   Dos pasos principales se realizan en el proceso de proyección:

   - Primero un subconjunto de puntos en S, llamado "puntos control" son proyectados en R^d por MDS (*Multidimensional Scaling*).
   - Haciendo uso de la relación de vecindad de los puntos en R^m y las coordenadas cartesianas de los puntos de control en R^d, es posible construir un sistema linear cuyas soluciones están en las coordenadas cartesianas de los puntos p_i en R^d

2. ***K-means :*** [6]
   Es el algoritmo más simple y más usado, aplicando un criterio de errores cuadrados [McQueen 1967]. Empieza con una partición aleatoria inicial y se mantiene reasignando los patrones a los clusters basado en la similitud entre patrones y los centros de los clusters hasta que el criterio de convergencia es alcanzado. Es un algoritmo popular por su fácil implementación y su complejidad es *O(n)*, donde *n* es el numero de patrones. El mayor problema de este algoritmo es que es sensible a la selección de la partición inicial y puede converger a un mínimo local si es que la partición inicial no fue escogida apropiadamente.

3. ***Principal Component Analysis (PCA) :***
   En la literatura de agrupamiento, PCA es a veces aplicada para reducir la dimensionalidad del conjunto de datos antes de agrupar. La idea de usar PCA antes de agrupar, es que PCA puede extraer la estructura del cluster en el conjunto de datos.[Jollife et al. 1980]. Según las conclusiones de [8]: "La calidad de los resultados de agrupamiento después de aplicar PCA no es necesariamente más alta que con sólo los datos originales."

4. ***Self Organizing Map (SOM) :*** [12]
   Una SOM consiste usualmente de un red de unidades, de 2 dimensiones. Cada unidad *i*, esta representada por un vector prototipo $m_i = [m_{i1}, ...m_{id}]$, donde *d* es la dimensión del vector de entrada. Las unidades están conectadas a sus adyacentes a través de una relación de vecindad. Durante el entrenamiento, la SOM forma una red elástica que se pliega en la *nube* formada por los datos de entrada.

7. Resultados y Discusión

Las pruebas fueron desarrolladas y probadas en una computador convencional: Pentium IV, con procesador AMD Turion 64X2 de 1.81GHz y 1.93GB de RAM.

A seguir mostramos el cálculo de firmas genéticas y su posterior clasificación usando *Least Square Projection* de la herramienta PEx, anteriormente mencionada, con los siguiente parámetros:

- Técnica de Proyección : Least Square Projection.
- Tipo de Distancia : Euclidiana.
- Algoritmo de Clustering : K-means.
- Número de Vecinos : 2.

Los casos a tomar en cuenta son los siguientes, tal como se mostró en la sección 5, donde se describe como se formó cada uno de los archivos de prueba :

(i) Una muestra por cada especie.

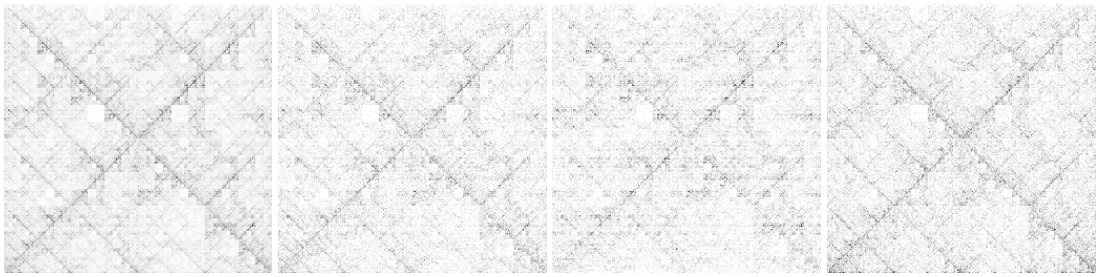
(ii) 50 muestras de tamaño 100 000, por cada especie.

(iii) Una muestra de tamaño 100 000 obtenida de las regiones codificantes, por cada especie.

(iv) Una muestra de tamaño 100 000 obtenida de las regiones no codificantes, por cada especie.

De forma ilustrativa, mostramos resultados para los organismos *A. fulgidus*, *E. coli* y V. cholerae. Los firmas genéticas obtenidas, aplicando FCGR, para cada uno de los casos anteriormente mencionados, se muestran en las figuras 4, 5 y 6. En estas imágenes se puede apreciar que la firma genética para cada uno de los casos es idéntica, pudiendo variar en los valores de las frecuencias de acuerdo al número de nucleótidos utilizados para realizar la firma.

En las figuras 7 y 8, se muestran los resultados de aplicar *Least Square Projection* a los archivos de prueba



(a) A. fulgidus-Genoma Completo (b) A. fulgidus-Subsecuencias (c) A. fulgidus-Exones (d) A. fulgidus-Introones

Figura 4: *A. fulgidus* - Firmas Genéticas de tamaño 2^8 obtenidas para cada uno de los casos [*(i),(ii),(iii),(iv)*].

(a) E. coli-Genoma Completo (b) E. coli-Subsecuencias (c) E. coli-Exones (d) E. coli-Introones

Figura 5: *E. coli* - Firmas Genéticas de tamaño 2^8 obtenidas para cada uno de los casos [*(i),(ii),(iii),(iv)*].

(a) V. cholerae-Genoma Completo (b) V. cholerae-Subsecuencias (c) V. cholerae-Exones (d) V. cholerae-Introones

Figura 6: *V. cholerae* - Firmas Genéticas de tamaño 2^8 obtenidas para cada uno de los casos [*(i),(ii),(iii),(iv)*].

para los casos *(i),(ii),(iii) y (iv)*, y se puede observar que en todos los casos se realiza un clustering apropiado, reconociendo siempre las 21 especies, además especies con proximidad filogenética se encuentran más cercanas.

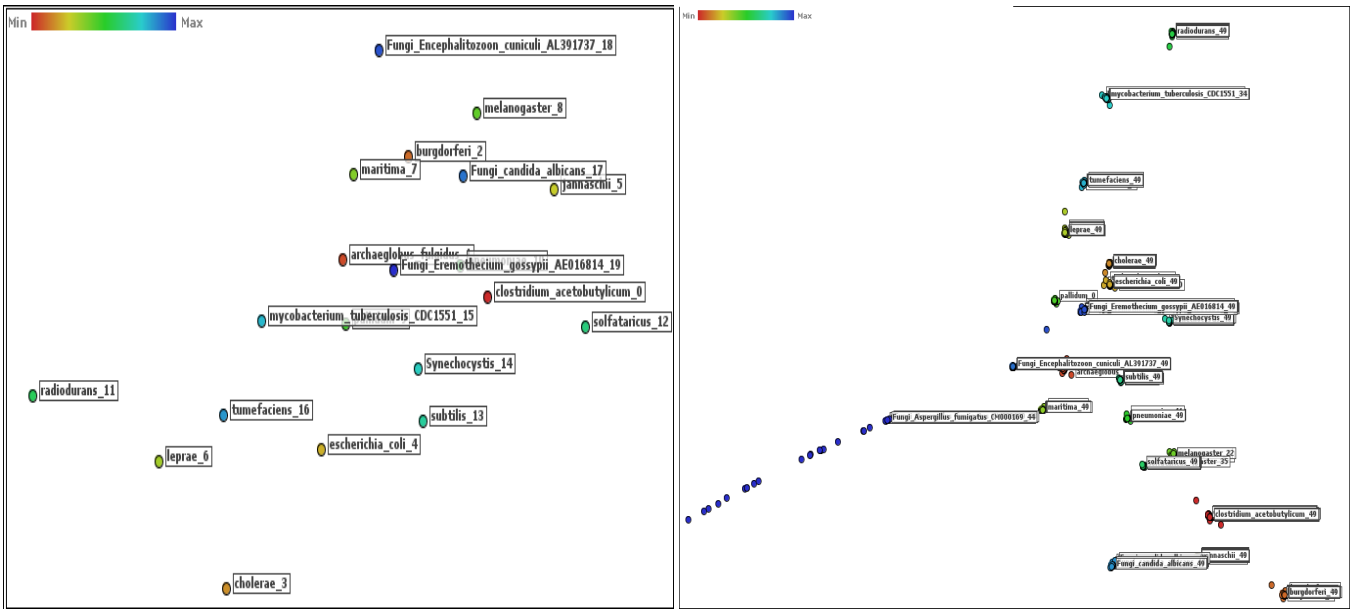
Por lo tanto, con nuestros experimentos realizados, acreditamos fuertemente que las firmas genéticas no tienen influencias de las regiones codificantes ni Regiones no Codificantes, las especies se pueden representar con tan sólo una subsecuencia del genoma completo, quedando abierta la posibilidad de evaluar el tamaño aproximado de oligonucleótidos para representar de una manera adecuada los genomas.

Este trabajo, correspondiente al tema de tesis del primer autor, aún continua en estudio, siendo que una de las tareas a realizar en un futuro inmediato es la clasificación de secuencias genómicas, usando firmas genéticas. En ese sentido, usamos algunas técnicas descritas en la sección 6, dando un mayor énfasis en el análisis de componentes principales y LSP.

En la figura 9, se puede apreciar los resultados obtenidos para PCA y LSP (usando la herramienta PEx) con archivos que contenían datos de 3 especies [10] [3] y SOM (usando Neural Network Toolbox de Matlab) con archivos que contenían 5 especies [4]

[3] PEx es una herramienta de visualización hecha en JAVA que puede ser usada para crear y explorar representaciones visuales de documentos y también puede ser usado para analizar otros tipos de datos multidimensionales. http://infoserver.lcad.icmc.usp.br/infovis2/PEx

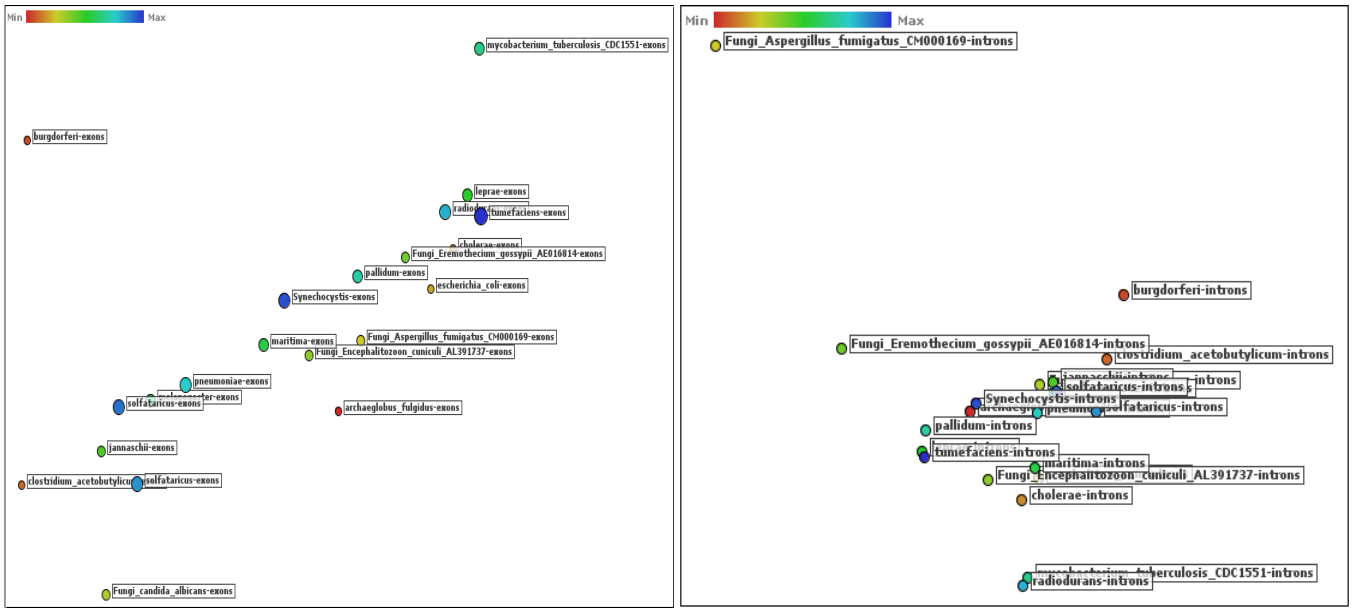
[4] Neural Network Toolbox : http://www.mathworks.com/products/neuralnet/



(a) *Least Square Projection* - Genoma Completo

(b) *Least Square Projection* - Subsecuencias

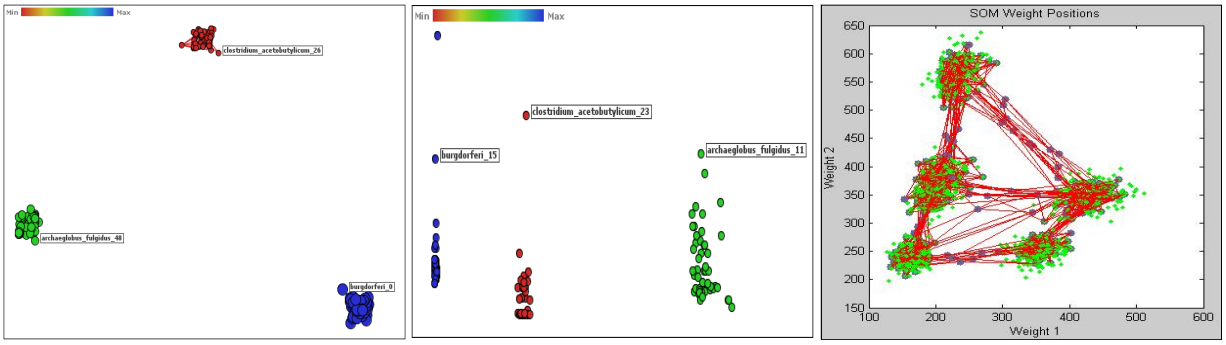
Figura 7: Clustering de Firmas Genéticas de tamaño 2^8 para los casos [*(i) y (ii)*].



(a) *Least Square Projection* - Exones

(b) *Least Square Projection* - Introones

Figura 8: Clustering de Firmas Genéticas de tamaño 2^8 para los casos [*(iii) y (iv)*].



(a) Least Square Projection (b) Principal Component Analysis (c) Self Organizing Map

Figura 9: Resultados para 3 técnicas de agrupamiento.

8. Conclusiones

- Comprobamos que aplicando FCGR a un genoma y sin la necesidad de aplicar PCA a los datos antes de realizar la clasificación, se forman 21 grupos diferentes, correspondientes a las especies utilizadas en las pruebas.

- También se comprobó que las firmas genéticas de subsecuencias de un genoma, son similares, permitiendo así la reducción de memoria y tiempo en el análisis de genomas.

- Así como se pudo apreciar en los resultados, preparatoriamente confirmamos el concepto de las firmas genéticas: *"especie-especifica"* pues siempre las especies se mostraban separadas a una distancia estadísticamente razonable, unas de otras.

- Se puede concluir que especies que mostraron firmas genéticas similares, fueron visualizadas en la imagen mucho más juntas que las demás.

- A partir de la imágenes obtenidas en los resultados, se ve que las firmas genéticas representan a las especies con tan sólo una subsecuencia del genoma, sin importar si pertenecen a las regiones codificantes o no codificantes.

- Una suposición a ser comprobada es el correspondiente al tamaño de oligonucleótidos a evaluar en el FCGR para obtener la firma genética, así como también la longitud de la subsecuencia del genoma a evaluar; de manera que se represente adecuadamente a las especies.

Referencias

[1] Maretzek Noble Fletcher Almeida, Carric. Analysis of genomic sequences by chaos game representation. 2001.

[2] A. Morris Ania L. Manson, Emma Jones. *Lo esencial en celula y genetica*, volume of . , , edition, . .

[3] A. Giron et al Deschavanne, P. Genomic signature: is preserved in short dna fragments. *BIBE 2000 IEEE international Symposium on bioinformatics biomedical engeneering , Washintown USA*, pages 161–167, november 2000.

[4] Vilain Fagot Fertil Deschavanne, Giron. Genomic signature:characterization and classification of species assessed by chaos game representation of sequences.

[5] Rosane Minghim Fernando V. Paulovich, Luis Gustavo Nonato and Haim Levkowitz. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, 14(3):565–566, MAY/JUNE 2008.

[6] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[7] H.J. Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Research.*, (18):2163–2170, 1990.

[8] W.L. Ruzzo K.Y. Yeung. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, Setiembre 2001.

[9] J. P. Mena-Chalco. Identificação de regiões codificantes de proteína através da transformada modificada de Morlet. Master's thesis, IME-USP, October 2005.

[10] Fernando V. Paulovich, Maria Cristina F. Oliveira, and Rosane Minghim. The projection explorer: A flexible tool for projection-based multidimensional visualization. In *Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing - SIBGRAPI*, pages 27–36, Belo Horizonte, Brazil, 2007. IEEE CS Press.

[11] Patrick DESCHAVANNE Sylvain LESPINATS, Alain GIRON and Bernard FERTIL. Dna sequences share a common syntax.

[12] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *Neural Networks, IEEE Transactions on*, 11(3):586–600, May 2000.

[13] Shiva Singh Yingwei Wang, Kathleen Hill and Lila Kari. The spectrum of genomic signatures: from dinucleotides to chaos game representation.